A Lasso for Hierarchical Interactions

Jacob Bien * Jonathan Taylor † Robert Tibshirani ‡

February 1, 2013

Abstract

We add a set of convex constraints to the Lasso to produce sparse interaction models that honor the hierarchy restriction that an interaction only be included in a model if one or both variables are marginally important. We give a precise characterization of the effect of this hierarchy constraint, prove that hierarchy holds with probability one, and derive an unbiased estimate for the degrees of freedom of our estimator. A bound on this estimate reveals the amount of fitting "saved" by the hierarchy constraint.

We distinguish between *parameter sparsity*—the number of nonzero coefficients—and *practical sparsity*—the number of raw variables one must *measure* to make a new prediction. Hierarchy focuses on the latter, which is more closely tied to important data collection concerns such as cost, time, and effort. We develop an algorithm, available in the `R` package `hierNet`, and perform an empirical study of our method.

Keywords: *Regularized regression, lasso, interactions, hierarchical sparsity, convexity*

1 Introduction.

There are numerous situations in which additive (main effects) models are insufficient for predicting an outcome of interest. In medical diagnosis, the co-occurrence of two symptoms may lead a doctor to be confident that a patient has a certain disease whereas the presence of either symptom without the other would provide only a moderate indication of that disease. This situation corresponds to a positive (i.e., synergistic) interaction between symptom variables. On the other hand, suppose both symptoms convey redundant information to the doctor about the patient so that knowing both provides no more information about the disease status than either one on its own. This situation is again not additive, but this time there is a negative interaction between symptoms. Fitting regression models with interactions is challenging when one has even a moderate number, p, of measured variables, since there are $\binom{p}{k}$ interactions of order k. For this paper, we focus on the case of pairwise ($k = 2$) interaction models, although the ideas we develop generalize naturally to higher-order interaction models.

1.1 Two-way interaction model.

We consider a regression model for an outcome variable Y and predictors $X_1, \ldots, X_p$, with pairwise interactions between these predictors. In particular our model has the form

$$Y = \beta_0 + \sum_j \beta_j X_j + \frac{1}{2}\sum_{j \neq k} \Theta_{jk} X_j X_k + \epsilon, \quad (1)$$

where $\epsilon \sim N(0,\ \sigma^2)$. Regardless of whether the predictors are continuous or discrete, we will refer to the additive part as the "main effect" terms and the quadratic part as the "interaction" terms. Our goal is to estimate $\beta \in \mathbb{R}^p$ and $\Theta \in \mathbb{R}^{p \times p}$, where $\Theta = \Theta^T$ and $\Theta_{jj} = 0$. The factor of one half before the interaction

*Department of Biological Statistics and Computational Biology and Department of Statistical Science, Cornell University, Ithaca, NY 14853; `jbien@cornell.edu`

†Department of Statistics, Stanford University, Stanford, CA 94305; `jonathan.taylor@stanford.edu`

‡Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305; `tibs@stanford.edu`

summation is a consequence of our notational decision to deal with a symmetric matrix Θ of interactions rather than a vector of length $p(p-1)/2$. We take $\Theta_{jj} = 0$ throughout this paper because it simplifies notation, but everything carries over if we remove this restriction and we provide this as an option in the `hierNet` (pronounced "hair net") package.

We observe a training sample, $(x_1, y_1), \ldots, (x_n, y_n)$, and our goal is to select a subset of the $p + p(p-1)/2$ main effect and interaction variables that is predictive of the response, and to estimate the values for the nonzero parameters of the model.

1.2 Strong and weak hierarchy.

It is a well-established practice among statisticians fitting (1) to only allow an interaction into the model if the corresponding main effects are also in the model. Such restrictions are known under various names, including "heredity," "marginality," and being "hierarchically well-formulated" (Hamada & Wu 1992, Chipman 1996, Nelder 1977, Peixoto 1987). There are two types of restrictions, which we will call *strong* and *weak hierarchy*:

$$\text{Strong Hierarchy:} \qquad \widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j \neq 0 \text{ and } \hat{\beta}_k \neq 0$$
$$\text{Weak Hierarchy:} \qquad \widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j \neq 0 \text{ or } \hat{\beta}_k \neq 0.$$

Some statisticians argue that models violating strong hierarchy are not sensible. For example, according to McCullagh & Nelder (1983),

> "[T]here is usually no reason to postulate a special position for the origin, so that the linear terms must be included with the cross-term."

To see that violating strong hierarchy amounts to "postulating a special position for the origin," consider writing an interaction model as $Y = \beta_0 + (\beta_1 + \Theta_{12}X_2)X_1 + \cdots$. First of all, we would only take $\beta_0 = 0$ if we have special reason to believe that the regression surface must go through the origin. Likewise, taking $\beta_1 = 0$ but $\Theta_{12} \neq 0$ would only be appropriate if we actually believe that X_1's effect on Y should only be present specifically when X_2 is nonzero. In most situations, we do not think that the variable X_2 that we measured is any more special than $aX_2 + b$. Yet if our model with X_2 violates strong hierarchy, then our model with $aX_2 + b$ (for any $b \neq 0$) is strongly hierarchical. This argument suggests that violations to hierarchy occur in special situations whereas hierarchy is the default.

Another argument in favor of hierarchy has to do with statistical power. In the words of Cox (1984):

> "[L]arge component main effects are more likely to lead to appreciable interactions than small components. Also, the interactions corresponding to larger main effects may be in some sense of more practical importance."

In other words, rather than looking at all possible interactions, it may be useful to focus our search to those interactions that have large main effects. Indeed, the method we propose in this paper makes direct use of this principle.

As a final argument for hierarchy, it is useful to distinguish between two notions of sparsity, which we will call *parameter sparsity* and *practical sparsity*. Parameter sparsity is what most statisticians mean by "sparsity": the number of nonzero coefficients in the model. Practical sparsity is what someone actually collecting data cares about: the number of variables one needs to *measure* to make predictions in the future. The hierarchy restriction favors models that "reuse" measured variables whereas a nonhierarchical model does not. The left panel of Figure 1 gives a small example where this difference is manifest. In fact, a simple calculation shows that this difference can be quite substantial: we can have a hierarchical and a nonhierarchical interaction model with the same parameter sparsity but with the nonhierarchical method having a practical sparsity of $k(k+1)$ where the hierarchical method's practical sparsity is just k.

While taking these arguments to the extreme leads to the use of strong hierarchy exclusively, we develop the case of weak hierarchy in parallel throughout this paper. Weak hierarchy, as the name suggests, can be thought of as a compromise between strong hierarchy and imposing no such structure and appears as a principle in certain statistical methods such as classification and regression trees (Breiman et al. 1984) and multivariate additive regression splines (Friedman 1991).

1.3 Sparsity, the Lasso, and Structured Sparsity.

The Lasso (Tibshirani 1996) is a method that performs both model selection and estimation. It penalizes the squared loss of the data with an ℓ_1-norm penalty on the parameter vector. This penalty has the property of producing estimates of the parameter vector that are sparse (corresponding to model selection). Given a design matrix $\widetilde{X} \in \mathbb{R}^{n\times d}$ and response vector $y \in \mathbb{R}^n$, the Lasso is the solution to the convex optimization problem:

$$\underset{\beta_0,\phi}{\text{Minimize}} \quad \frac{1}{2}\|y - \beta_0 1 - \widetilde{X}\phi\|^2 + \lambda\|\phi\|_1.$$

The penalty parameter, $\lambda \geq 0$, controls the relative importance of fitting to the training data (sum-of-squares term) and of sparsity (ℓ_1 penalty term). A natural extension of the Lasso to our interaction model (1) would be to take $\phi^T = [\beta^T, \mathrm{vec}(\Theta)^T]$ and $\widetilde{X} = (X : Z/2)$, where the columns of $Z \in \mathbb{R}^{n\times p(p-1)}$ correspond to elementwise products of the columns of X. We will refer to this method as the *All-Pairs Lasso* since it is simply the Lasso applied to a data matrix which includes all pairs of interactions (as well as all main effects). It is common with the Lasso to standardize the predictors so that they are on the same scale. In this paper, we standardize X so that its columns have mean 0 and standard deviation 1; we then form Z from these standardized predictors, and finally center the resulting columns of Z. By centering y and $\widetilde{X}$, we may take $\hat{\beta}_0 = 0$.

The Lasso's ℓ_1 penalty is neutral to the pattern of sparsity, allowing any sparsity pattern to emerge. The notions of strong and weak hierarchy introduced in Section 1.2 represent situations in which we want to exclude certain sparsity patterns. There has been a growing literature focusing on methods that produce *structured sparsity* (Yuan & Lin 2006, Zhao et al. 2009, Jenatton et al. 2010, 2011, Bach 2011, Bach et al. 2012). These methods make use of the *group lasso* penalty (and generalizations thereof), which, given a predetermined grouping of the parameters, induce entire groups of parameters to be set to zero (Yuan & Lin 2006). Given a set of groups of variables, $\mathcal{G}$, these methods generalize the ℓ_1 penalty by

$$\sum_{G\in\mathcal{G}} d_G\|\phi_G\|_{\gamma_G},$$

where $\gamma_G > 1$, ϕ_G is ϕ projected onto the coordinates in G, and d_G is a nonnegative weight. Hierarchical structured sparsity is obtained by choosing $\mathcal{G}$ to have nested groups. For example, Zhao et al. (2009) consider the penalty

$$\sum_{j\neq k} \left\{|\Theta_{jk}| + \|(\beta_j, \beta_k, \Theta_{jk})\|_{\gamma_{jk}}\right\}.$$

Likewise, the framework of Bach et al. (2012) if specialized to this paper's focus would lead to a penalty of the form

$$\|\Theta\|_1 + \sum_j d_j\|(\Theta_j, \beta_j)\|_q \tag{2}$$

for some $q > 1$ and $d_j > 0$. In fact, Radchenko & James (2010) suggest a penalty for generalized additive models with interactions that reduces to (2) in the linear model case, with $q = 2$ and d_j independent of j.

1.4 This paper.

Here, we propose a lasso-like procedure that produces sparse estimates of β and Θ while satisfying the strong or weak hierarchy constraint (Section 2). In contrast to much of the structured sparsity literature which is based on group lasso penalties, our approach, presented in Section 2, involves adding a set of convex constraints to the Lasso. Although we find this form of constraint more naturally interpretable, we show (Remark 3) that this problem can be equivalently expressed in a form that relates it to penalties from the structured sparsity literature such as (2).

A key advantage of our specific choice of penalty structure is that it admits a simple interpretation of the effect of the hierarchy demand. Unlike other hierarchical sparsity methods, which do not pay much attention

to the particular choice of norms (as long as $\gamma_G > 1$), our formulation is carefully tailored to allow it to be related directly back to the Lasso, permitting one to understand specifically how hierarchy alters the solution (Section 3.1). This feature of our estimator gives it a transparency that exposes the effects (both positive and negative!) of the hierarchy constraint. Furthermore, our characterization suggests that the demand for hierarchy is—analogous to the demand for sparsity—a form of "regularization." We develop an unbiased estimator of the degrees of freedom of our method (Section 3.3) and an interpretable upper bound on this quantity, which also points to hierarchy as regularization. In particular, we show that we do not "spend" in degrees of freedom for main effects that are forced into the model by the hierarchy constraint.

Another difference from much of the structured sparsity literature, which aims to develop a broad treatment of structured and hierarchical sparsity methods, is that our focus is narrowed to the problem of interaction models. Our restricted scope allows us to address specifically the performance of such a tool to this important problem. In Section 4, we review previous work on the problem of hierarchical interaction model fitting and selection. These methods fall into three categories: Multi-step procedures, which are defined by an algorithm (Peixoto 1987, Friedman 1991, Turlach 2004, Nardi & Rinaldo 2007, Bickel et al. 2008, Park & Hastie 2008, Wu et al. 2010); Bayesian approaches, which specify the hierarchy requirement through a prior (Chipman 1996); and, most related to this paper's proposal, regularized regression methods, which are defined by an optimization problem (Yuan et al. 2009, Zhao et al. 2009, Choi et al. 2010, Jenatton et al. 2010, Radchenko & James 2010). In Section 5, we study via simulation the statistical implications of imposing hierarchy on an interactions-based estimator under various scenarios (in both the Lasso and stepwise frameworks). In Section 6 we present an efficient algorithm for computing our estimator. Real data examples are used to illustrate a distinction we draw between "parameter sparsity" and "practical sparsity" and to discuss hierarchy's role in promoting the latter.

2 Our proposed method.

In Section 1.3, we introduced the *All-Pairs Lasso*, which can be written as

$$\underset{\beta_0\in\mathbb{R},\ \beta\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad q(\beta_0,\beta,\Theta)+\lambda\|\beta\|_1+\frac{\lambda}{2}\|\Theta\|_1 \quad \text{s.t.} \quad \Theta=\Theta^T, \tag{3}$$

where $\|\Theta\|_1=\sum_{j\neq k}|\Theta_{jk}|$ and $q(\beta_0,\beta,\Theta)$ is the loss function, typically $\frac{1}{2}\sum_{i=1}^n(y_i-\beta_0-x_i^T\beta-\frac{1}{2}x_i^T\Theta x_i)^2=\frac{1}{2}\|y-\beta_0 1-X\beta-Z\text{vec}(\Theta)/2\|^2$, but may also include a ridge penalty on the coefficients as discussed later or may be substituted for the binomial negative log-likelihood. The one-half factors in front of terms involving Θ are merely a consequence of the notational choice to represent Θ as a symmetric matrix (with $\Theta_{jj}=0$ for $j=1,\ldots,p$). In this paper, we propose a modification of the All-Pairs Lasso that produces models that are guaranteed to be hierarchical.

As motivation for our proposal, consider building hierarchy into the optimization problem as a constraint:

$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta)+\lambda\|\beta\|_1+\frac{\lambda}{2}\|\Theta\|_1 \\ & \text{s.t.} \quad \Theta=\Theta^T, \quad \|\Theta_j\|_1\le|\beta_j| \text{ for } j=1,\ldots,p,\end{aligned} \tag{4}$$

where Θ_j denotes the jth row (and column, by symmetry) of Θ. Notice that if $\widehat{\Theta}_{jk}\neq 0$, then $\|\widehat{\Theta}_j\|_1>0$ and $\|\widehat{\Theta}_k\|_1>0$ and thus $\hat{\beta}_j\neq 0$ and $\hat{\beta}_k\neq 0$. While the added constraints enforce strong hierarchy, they are not convex, which makes (4) undesirable as a method. In this paper, we propose a straightforward convex relaxation of (4), which we call the *Strong Hierarchical Lasso*:

$$\begin{aligned}\underset{\substack{\beta_0\in\mathbb{R},\ \beta^\pm\in\mathbb{R}^p,\\ \Theta\in\mathbb{R}^{p\times p}}}{\text{Minimize}} \quad & q(\beta_0,\beta^+-\beta^-,\Theta)+\lambda 1^T(\beta^++\beta^-)+\frac{\lambda}{2}\|\Theta\|_1 \\ & \text{s.t.} \quad \Theta=\Theta^T, \quad \left.\begin{aligned}\|\Theta_j\|_1\le\beta_j^++\beta_j^-\\ \beta_j^+\ge 0,\beta_j^-\ge 0\end{aligned}\right\}\text{for } j=1,\ldots,p,\end{aligned} \tag{5}$$

where we have replaced the optimization variable $\beta \in \mathbb{R}^p$ by two vectors $\beta^+, \beta^- \in \mathbb{R}^p$. After solving the above problem, our fitted model is of the form $\hat{f}(x) = \hat{\beta}_0 + x^T(\hat{\beta}^+ - \hat{\beta}^-) + x^T\widehat{\Theta}x/2$. While we might informally think of β^+ and β^- as positive and negative parts of a vector $\beta = \beta^+ - \beta^-$, i.e. that $\beta^\pm = \max\{\pm\beta, 0\}$, this is not actually the case since at a solution we can have both $\hat{\beta}_j^+ > 0$ and $\hat{\beta}_j^- > 0$. Indeed, if we were to add the constraints $\beta_j^+\beta_j^- = 0$ for $j = 1, \ldots, p$ to (5), then these would be positive and negative parts and so $\beta_j^+ + \beta_j^- = |\beta_j|$, giving us precisely problem (4). This observation establishes that (5) is a convex relaxation of (4).

The hierarchy constraints can be seen as an embedding into our method of David Cox's "principle" that "large component main effects are more likely to lead to appreciable interactions than small components." The constraint

$$\|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-$$

budgets the total amount of interactions involving variable X_j according to the relative importance of X_j as a main effect. One additional advantage of the convex relaxation is that the constraint is less restrictive. If the best fitting model would have $\|\Theta_j\|_1$ large but $|\beta_j|$ only moderate, this can be accommodated by making β_j^+ and β_j^- both large.

Remark 1: Another possibility for the hierarchy constraint that we have considered is $|\Theta_{jk}| \le \beta_j^+ + \beta_j^-$; however, we have found that this can lead to an overabundance of interactions relative to main effects.

Remark 2: It is desirable to include in the loss function q an *elastic net* term, $(\epsilon/2)\left(\|\Theta\|_F^2 + \|\beta^+\|^2 + \|\beta^-\|^2\right)$, to ensure uniqueness of the solution (Zou & Hastie 2005). We think of $\epsilon > 0$ as a fixed tiny fraction of λ, such as $\epsilon = 10^{-8}\lambda$, rather than as an additional tuning parameter. Such a modification does not complicate the algorithm, but simplifies the study of the estimator. In all numerical examples and in the `hierNet` package, we use this elastic net modification.

Remark 3: We prove in Appendix A that (5) may equivalently be written as

$$\underset{\substack{\beta_0\in\mathbb{R},\ \beta\in\mathbb{R}^p,\\ \Theta\in\mathbb{R}^{p\times p}, \Theta=\Theta^T}}{\text{Minimize}} \quad q(\beta_0, \beta, \Theta) + \lambda\sum_j \max\{|\beta_j|, \|\Theta_j\|_1\} + \frac{\lambda}{2}\|\Theta\|_1. \tag{6}$$

This reparameterization of the problem shows its similarities to the group lasso based methods. In place of the more standard penalty $\|(\Theta_j, \beta_j)\|_q$ of (2), we use $\max\{\|\Theta_j\|_1, |\beta_j|\}$. In Section 3.1, we show that this unusual choice of penalty admits a particularly simple interpretation for the effect of imposing hierarchy.

In Section 1.2, we also introduced the notion of weak hierarchy. By simply removing the symmetry constraint on Θ, we get what we call the *Weak Hierarchical Lasso*:

$$\begin{aligned} \underset{\beta_0\in\mathbb{R},\ \beta^\pm\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad & \left.\begin{aligned} \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^- \\ \beta_j^+ \ge 0, \beta_j^- \ge 0 \end{aligned}\right\} \text{for } j = 1, \ldots, p. \end{aligned} \tag{7}$$

Even though at a solution to this problem, $\widehat{\Theta}$ is not symmetric, we should think of the interaction coefficient as $(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2$ since this is what multiplies the interaction term $x_{ij}x_{ik}$ when computing $\hat{f}(x_i)$.

Remark 4: We can build further on the connection between (2) and (5) discussed in Remark 3. Our removal of the symmetry constraint in (7) is analogous to the technique of duplicating columns of the design matrix used in the overlap group lasso (Obozinski et al. 2011).

A favorable property that distinguishes our method from previous approaches discussed in Section 4 is the relative transparency of the role that the hierarchy constraint plays in our estimator. This aspect is developed in Section 3.1.

Although our primary focus in this paper is on the Gaussian setting of (1), our proposal extends straightforwardly to other situations, such as the logistic regression setting in which the response is binary. In this case, we simply have $q(\beta_0, \beta, \Theta)$ be the appropriate negative log-likelihood: $-\sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i)$, where $p_i = [1 + \exp(-\beta_0 - x_i^T\beta - \frac{1}{2}x_i^T\Theta x_i)]^{-1}$. In Appendix F, we show that to solve this problem requires only a minor modification to our primary algorithm. It should also be noted that our estimator (and the algorithms developed to compute it) is designed for both the $p < n$ and $p \geq n$ setting.

As a preliminary example, consider predicting whether a sample of olive oil comes from Southern Apulia based on measurements of the concentration of $p = 8$ fatty acids (Forina et al. 1983). The dataset consists of $n = 572$ samples, and we average our results over 100 random equal-sized train-test splits. We compare three methods: (a) a standard Lasso with main effects only (MEL), (b) the All-Pairs Lasso (APL), and (c) the Strong Hierarchical Lasso (HL).

The top left panel of Figure 1 shows an interesting difference between HL and APL. We see that, on average, at a parameter sparsity level of five, the HL model uses four of the measured variables whereas APL uses six. Using the hierarchical model to classify a future olive oil, we only need to measure four rather than six of the fatty acids.

The top right panel of Figure 1 shows the predictive performance (versus the practical sparsity) of the three methods. It appears that HL enjoys the "best of both worlds," matching the good performance of MEL for low practical sparsity levels (since it tends to pick out the main effects first) and the good performance of APL at high practical sparsity levels (since it can incorporate predictive interactions). Finally, the bottom panel of the figure provides a visual display of a sequence of HL's solutions (by varying λ). Nonzero main effects are shown as filled nodes and edges indicate nonzero interactions. Since all edges are incident to filled nodes, we see that strong hierarchy holds.

In the next section, we present several properties of our estimator that shed light on the effect of adding the convex hierarchy constraint to the Lasso. Among these properties is an unbiased estimate of the degrees of freedom of our estimator. We view this degrees of freedom result as valuable primarily for the sake of understanding the effect of hierarchy. While such an estimate could be used for parameter selection, we prefer cross validation to select λ since this is more directly tied to the goal of prediction.

3 Properties.

3.1 Effect of the constraint.

A key advantage of formulating an estimator as a solution to a convex problem is that it can be completely characterized by a set of optimality conditions, known as the Karush-Kuhn-Tucker (KKT) conditions. These conditions are useful for understanding the effect that the hierarchy constraint in (5) and (7) has on our solutions. In this section, we will study the simplest case, taking $q(\beta_0, \beta, \Theta)$ to be the quadratic loss function with no elastic net penalty. We let

$$\begin{aligned} r^{(-j)} &= y - \hat{y} + x_j\hat{\beta}_j \\ r^{(-jk)} &= y - \hat{y} + (x_j * x_k)(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2 \end{aligned}$$

denote partial residuals (where $*$ denotes elementwise multiplication), and we assume that $\|x_j\|_2 = 1$. For linear regression, the KKT conditions are known as the normal equations and can be written as

$$\hat{\beta}_j = x_j^T r^{(-j)} \qquad\qquad \widehat{\Theta}_{jk} = \frac{(x_j * x_k)^T r^{(-jk)}}{\|x_j * x_k\|^2}.$$

The All-Pairs Lasso solution satisfies

$$\hat{\beta}_j = \mathcal{S}(x_j^T r^{(-j)},\ \lambda) \qquad\qquad \widehat{\Theta}_{jk} = \frac{\mathcal{S}[(x_j * x_k)^T r^{(-jk)},\ \lambda]}{\|x_j * x_k\|^2}. \tag{8}$$

where $\mathcal{S}$ denotes the soft-thresholding operator defined by $\mathcal{S}(c,\ \lambda) = \text{sign}(c)(|c| - \lambda)_+$. Written this way, we see that the Lasso is similar to linear regression, but all coefficients are shrunken towards 0, with some

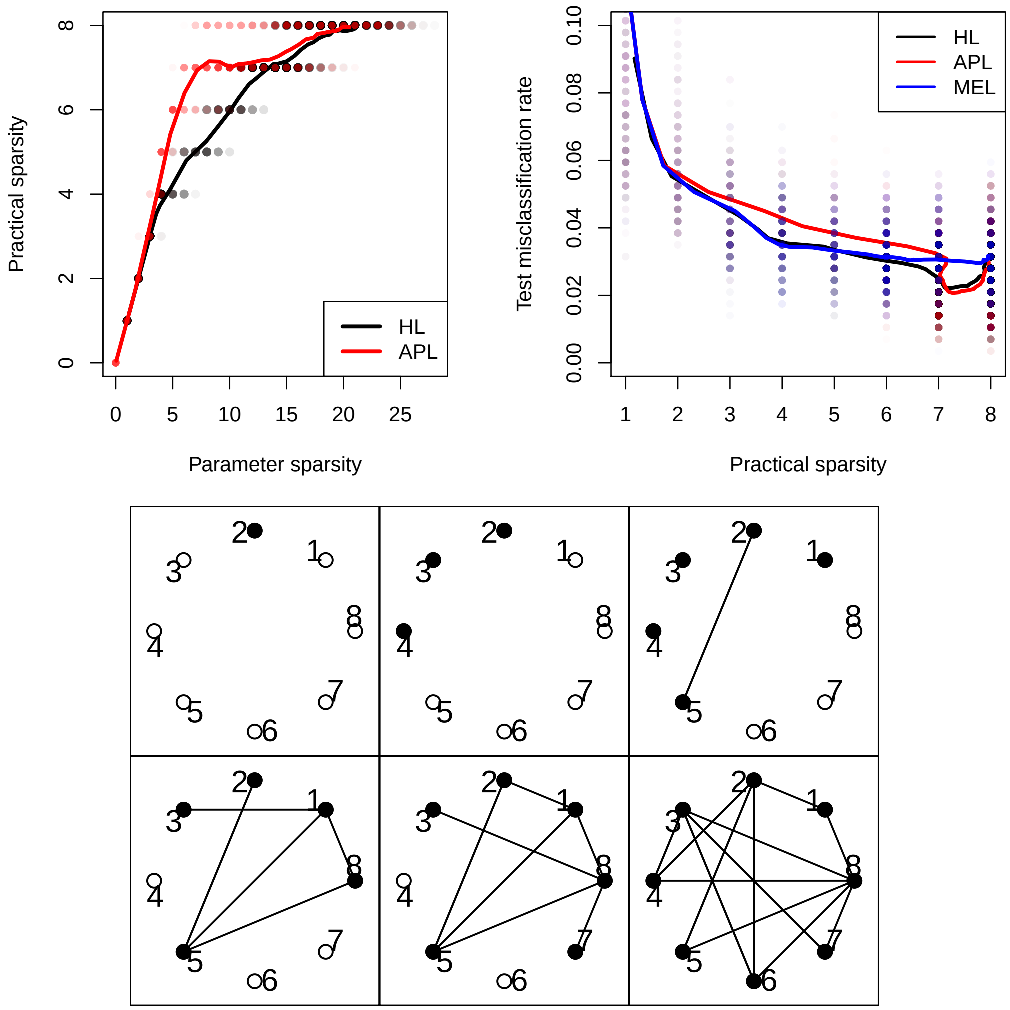

Figure 1: Olive oil data: (Top left) Parameter sparsity is the number of nonzero coefficients while practical sparsity is the number of *measured* variables in the model. Results from all 100 random train-test splits are shown as points; lines show the average performance over all 100 runs. (Top right) Misclassification error on test set versus practical sparsity. (Bottom) Wheel plots showing the sparsity pattern at 6 values of λ for the Strong Hierarchical Lasso. Filled nodes correspond to nonzero main effects and edges correspond to nonzero interactions.

coefficients (those for which $|x_j^T r^{(-j)}| \leq 0$) set to zero. It is instructive to examine the corresponding statements for the Strong and Weak Hierarchical Lasso methods.

Property 1. *The coefficients of the Strong and Weak Hierarchical Lassos with $\lambda > 0$ and taking $q(\beta_0, \beta, \Theta)$ to be the quadratic loss (with no elastic net penalty) satisfy*

- Strong:

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda - \hat{\alpha}_j)$$
$$\widehat{\Theta}_{jk} = \frac{\mathcal{S}[(x_j * x_k)^T r^{(-jk)},\ \lambda + \hat{\alpha}_j + \hat{\alpha}_k]}{\|x_j * x_k\|^2},$$

- Weak:

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda - \tilde{\alpha}_j)$$
$$\frac{\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj}}{2} = \frac{\mathcal{S}[(x_j * x_k)^T r^{(-jk)},\ \lambda + 2\min\{\tilde{\alpha}_j, \tilde{\alpha}_k\}]}{\|x_j * x_k\|^2},$$

for some $\hat{\alpha}_j \geq 0$, $j = 1, \ldots, p$ with $\hat{\alpha}_j = 0$ when $\|\widehat{\Theta}_j\|_1 < \hat{\beta}_j^+ + \hat{\beta}_j^-$ (and likewise for $\tilde{\alpha}_j$).

Proof. See Appendix B. □

The $\hat{\alpha}_j, \tilde{\alpha}_j$ appearing in the above two properties are optimal dual variables corresponding to the jth hierarchy constraint for the Strong and Weak Hierarchical Lasso problems, respectively. When $\|\widehat{\Theta}_j\|_1 < \hat{\beta}_j^+ + \hat{\beta}_j^-$, we have $\hat{\alpha}_j = 0$ (or $\tilde{\alpha}_j = 0$) by complementary slackness. Comparing these expressions to those of the All-Pairs Lasso gives insight into the effect of the constraint. Property 1 reveals that the overall form of the All-Pairs Lasso and Hierarchical Lasso methods are identical. The difference is that the hierarchy constraint leads to a reduction in the shrinkage of certain main effects and an increase in the shrinkage of certain interactions. In particular, we see that when the hierarchy constraints are loose at the solution, i.e. $\|\widehat{\Theta}_j\|_1 < \hat{\beta}_j^+ + \hat{\beta}_j^-$, the Weak Hierarchical Lasso's optimality conditions become identical to the All-Pairs Lasso (since $\tilde{\alpha}_j = 0$) for all coefficients involving x_j. For the Strong Hierarchical Lasso, when both the jth and kth constraints are loose, the optimality conditions match those of the All-Pairs Lasso for the coefficients of x_j, x_k, and $x_j * x_k$. The methods differ when constraints are active, i.e. when $\|\widehat{\Theta}_j\|_1 = \hat{\beta}_j^+ + \hat{\beta}_j^-$, which allows $\hat{\alpha}_j$ (or $\tilde{\alpha}_j$) to be nonzero. Intuitively, this case corresponds to the situation in which hierarchy would have not held "naturally" (i.e. without the constraint) and the corresponding dual variable plays the role of reducing $\widehat{\Theta}_j$ in ℓ_1-norm and increasing $\hat{\beta}_j^+ + \hat{\beta}_j^-$ until the constraint is satisfied. The way in which the Weak and Strong Hierarchical Lasso methods perform this shrinkage is different, but both are identical to All-Pairs Lasso when all constraints are loose.

3.2 Hierarchy guarantee.

In Section 2, we showed that adding the constraint $\|\Theta_j\|_1 \leq |\beta_j|$ would guarantee that hierarchy holds. However, we have not yet shown that the same is true of the convex relaxation's constraint, $\|\Theta_j\|_1 \leq \beta_j^+ + \beta_j^-$. In particular, while $\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ + \hat{\beta}_j^- \neq 0$, we could still have $\hat{\beta}_j^+ - \hat{\beta}_j^- = 0$. This would correspond to a model in which $X_j X_k$ is used in the model, but X_j is not. Intuitively, we would expect that if $\hat{\beta}_j^+ > 0$, then $\hat{\beta}_j^+ = \hat{\beta}_j^-$ is analogous to getting an exact zero in linear regression (i.e. a zero probability event). In this section, we establish that this is in fact the case.

In particular, we study (5) and (7) where $q(\beta_0, \beta, \Theta)$ includes an elastic net term. The importance of this modification is that it ensures uniqueness, simplifying the analysis. As noted in Remark 2, we think of ϵ as a small, fixed proportion of λ rather than as a separate tuning parameter.

Property 2. *Suppose y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$. If $(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})$ solves* (5)*, where $q(\beta_0, \beta, \Theta)$ is the quadratic loss with an $\epsilon > 0$ ridge penalty, then strong hierarchy holds with probability 1, i.e.*

$$\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0 \text{ AND } \hat{\beta}_k^+ - \hat{\beta}_k^- \neq 0.$$

Proof. See Appendix C. $\square$

To understand how dropping the symmetry constraint leads to the "or" statement required of weak hierarchy, note that $X_j X_k$ is in the Weak Hierarchical Lasso model if and only if $\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj} \neq 0$. This holds only if $\widehat{\Theta}_{jk} \neq 0$ or $\widehat{\Theta}_{kj} \neq 0$.

Property 3. *Suppose y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$. If $(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})$ solves (7), where $q(\beta_0, \beta, \Theta)$ is the quadratic loss with an $\epsilon > 0$ ridge penalty, then weak hierarchy holds with probability 1, i.e.*

$$(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2 \neq 0 \implies \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0 \text{ OR } \hat{\beta}_k^+ - \hat{\beta}_k^- \neq 0.$$

Proof. See Appendix C. $\square$

3.3 Degrees of Freedom.

In classical statistics, the degrees of freedom of a procedure refer to the dimension of the space over which its fitted values can vary. It is useful in that it provides a measure of how much "fitting" the procedure is doing. This notion can be generalized to adaptive procedures such as the Lasso (Stein 1981, Efron 1986, Efron et al. 2004, Zou et al. 2007). See (Ryan) Tibshirani & Taylor (2012) for a thorough discussion. If given data $y \in \mathbb{R}^n$, a procedure h produces fitted values $\hat{y} = h(y) \in \mathbb{R}^n$, the degrees of freedom of the procedure h is defined to be

$$df(h) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{y}_i). \tag{9}$$

Property 4. *Suppose $y \sim N(\mu, \sigma^2 I_n)$. An unbiased estimate of the degrees of freedom of the Strong Hierarchical Lasso, with quadratic loss and no ridge penalty, is given by*

$$\widehat{df}_\lambda = \mathrm{rank}(\widetilde{X} P),$$

where $\widetilde{X} = (X : -X : Z/2 : -Z/2)$ with Z containing the interactions, and P is a projection matrix which depends on the sign pattern of $(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})$ and on the set of hierarchy constraints that are tight.

Proof. See Appendix D. $\square$

Figure 2 provides a numerical evaluation of how well $\widehat{\mathrm{df}}_\lambda$ estimates df_λ. We fix $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, and $\Theta \in \mathbb{R}^{p \times p}$, and we generate $B = 10,000$ Monte Carlo replicates $y^{(1)}, \ldots, y^{(B)} \in \mathbb{R}^n$. For each replicate, we fit the Strong Hierarchical Lasso along a grid of λ values to get $(\hat{\beta}_\lambda^{+(b)}, \hat{\beta}_\lambda^{-(b)}, \widehat{\Theta}_\lambda^{(b)})$ and $\hat{y}_\lambda^{(b)} \in \mathbb{R}^n$. From these values, we compute Monte Carlo estimates of df_λ from the definition in (9) and of $E[\widehat{\mathrm{df}}_\lambda]$.

While $\widehat{df}_\lambda$ can be calculated from the data and is therefore useful as an unbiased way of calibrating the amount of fitting the Strong Hierarchical Lasso is doing, this expression is difficult to interpret. However, it turns out that we can *bound* $\widehat{df}_\lambda$ by a quantity that does make more sense.

Property 5. *Let $\mathcal{T} = \{j : \|\widehat{\Theta}_j\|_1 = \hat{\beta}_j^+ + \hat{\beta}_j^-\}$, $\mathcal{A}_\beta = \{j : \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0\}$, $\mathcal{A}_\pm = \{j : \hat{\beta}_j^\pm > 0\}$, and $\mathcal{A}_\Theta = \{jk : \widehat{\Theta}_{jk} \neq 0, j < k\}$. Then,*

$$\widehat{df}_\lambda \leq |\mathcal{A}_\beta| + |\mathcal{A}_\Theta| - |\mathcal{T} \cap (\mathcal{A}_+ \Delta \mathcal{A}_-)|$$

holds almost surely, where $\mathcal{A}_+ \Delta \mathcal{A}_- = (\mathcal{A}_+ \setminus \mathcal{A}_-) \cup (\mathcal{A}_- \setminus \mathcal{A}_+)$.

Proof. See Appendix D. $\square$

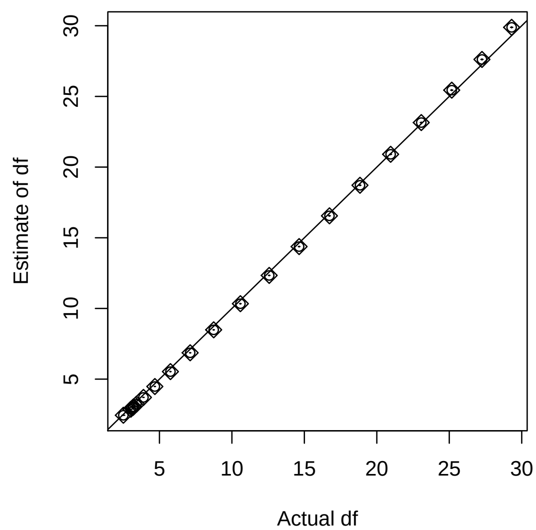

Figure 2: Numerical evaluation of how well $\widehat{\mathrm{df}}_\lambda$ estimates df_λ. Monte Carlo estimates of $E[\widehat{\mathrm{df}}_\lambda]$ (y-axis) versus Monte Carlo estimates of df_λ (x-axis) for a sequence of λ values (circular) are shown. One standard error bars are drawn and are hardly visible. Our bound on the unbiased estimate is plotted with diamonds.

By contrast, for the All-Pairs Lasso in the case that $p + \binom{p}{2} < n$ and the design matrix is full rank, we have $df_\lambda(\mathrm{APL}) = E[|\mathcal{A}_\beta| + |\mathcal{A}_\Theta|]$ (Zou et al. 2007). In other words, the Strong Hierarchical Lasso does not "pay" (in terms of fitting) for those main effects, $\hat{\beta}_j^+ - \hat{\beta}_j^-$, that are forced into the model by the hierarchy constraint to accommodate a strong interaction. Notice that we do pay for a nonzero main effect if *both* $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ are nonzero. This makes sense since the constraint could be satisfied with just one of these variables nonzero, but in this case it is advantageous to the fit to make both nonzero. In Figure 2, we find that this bound is in expectation visually indistinguishable from $E[\widehat{df}_\lambda]$.

4 Related Work.

There has been considerable interest in fitting interaction models in statistics and related fields. We focus here on an overview of methods that aim at forming predictive models that satisfy the hierarchical interactions restriction.

4.1 Multi-step procedures.

Many statistics textbooks discuss a simple stepwise procedure in which one iteratively considers adding or removing the "best" variable (whether it be main effect or interaction); they add that one should only consider including an interaction if its main effects are in the model (e.g., see backward elimination in Agresti 2002, Section 6.1.3). In doing so, they are enforcing the strong hierarchy restriction. Such procedures are ubiquitous (Nelder 1997, Peixoto 1987) as are more recent versions (Friedman 1991, Bickel et al. 2008, Park & Hastie 2008, Wu et al. 2010). Another approach is to perform model selection first without considering hierarchy and then to include any lower-order terms necessary to satisfy hierarchy as a post-processing step (Nardi & Rinaldo 2007). Finally, Turlach (2004) and Yuan et al. (2007) consider modifying the LARS algorithm (Efron et al. 2004) so that hierarchy is enforced.

4.2 Bayesian approaches.

Another set of procedures for building hierarchical interaction models comes from a Bayesian viewpoint. Chipman (1996) adapt the *Stochastic Search Variable Selection* (SSVS) approach of George & McCulloch (1993) to produce strong or weak hierarchical interaction models. SSVS makes use of a hierarchical normal

mixture model to perform variable selection in regression. Every variable has a latent binary variable indicating whether it is "active." Conditional on this latent variable, each coefficient is a 0-mean normal with variance determined by the latent importance of the coefficient. The original SSVS paper chooses a prior in which the importance of each variable is an independent Bernoulli. Chipman (1996) introduces dependence into the prior so that Θ_{jk} is important only if β_j and/or β_k is as well.

4.3 Optimization-based approaches.

Choi et al. (2010) formulate a nonconvex optimization problem to get sparse hierarchical interaction models. They write $\Theta_{jk} = \Gamma_{jk}\beta_j\beta_k$, where β are the main effect coefficients and then apply ℓ_1 penalties on β and Γ. Notice that $\Theta_{jk} \neq 0$ implies $\beta_j \neq 0$ and $\beta_k \neq 0$. The nonconvexity arises in writing Θ_{jk} as the product of optimization variables.

Most similar to this paper's proposal are a series of methods which formulate convex optimization problems to give sparse hierarchical interaction models. Yuan et al. (2009) modify the nonnegative garrote (Breiman 1995) by adding linear inequality constraints to enforce hierarchy. In this sense, our method can be seen as the adaption of their approach to the Lasso.

Finally, as discussed in Section 1.3, another set of convex methods make use of the Group Lasso penalty (Yuan & Lin 2006). Zhao et al. (2009) (and, relatedly, Jenatton et al. 2011) describe Composite Absolute Penalties (CAP), a very broad class of penalties that can achieve group and hierarchical sparsity. To achieve "hierarchical selection," they put forward the principle that a penalty of the form $\|(\phi_1, \phi_2)\|_\gamma + |\phi_1|$, with $\gamma > 1$, induces ϕ_2 to be zero only when ϕ_1 is zero as well. For hierarchical interaction models, they suggest a penalty of the form $\lambda \sum_{j<k} \left[|\Theta_{jk}| + \|(\beta_j, \beta_k, \Theta_{jk})\|_{\gamma_{j,k}}\right]$. This framework has been developed in the structured sparsity literature (e.g., Bach et al. 2012). Radchenko & James (2010) introduce VANISH, which uses this nested-group principle to achieve hierarchical sparsity in the context of non-linear interactions. Their penalty in the setting of (1) is $\sum_j [\lambda_1 \|(\beta_j, \Theta_j)\|_2 + \lambda_2 \|\Theta_j\|_1]$. As noted in Remark 3, our proposal is closer to CAP and VANISH than it may first appear. Our problem can be rewritten to have a penalty of the form $\lambda \sum_j [\max\{|\beta_j|, \|\Theta_j\|_1\} + (1/2)\|\Theta_j\|_1]$. In this sense, the penalty is in the spirit of CAP and related methods, although it does not quite fall into the class of CAP (since ours involves a sum of norms of norms). It is most similar to VANISH in that it combines all of Θ_j into the term involving β_j.

5 Empirical Study.

5.1 Simulations.

Our main interest in this section is to study the advantages and disadvantages of restricting one's interaction models to those that honor hierarchy. Clearly, the effectiveness of such a strategy depends on the true model generating the data. We take $n = 100$ and $p = 30$ (435 two-way interactions) and consider four scenarios:

(I) Truth is hierarchical: $\Theta_{jk} \neq 0 \implies \beta_j \neq 0,\ \beta_k \neq 0$.

(II) Truth is anti-hierarchical: $\Theta_{jk} \neq 0 \implies \beta_j = 0,\ \beta_k = 0$.

(III) Truth only has interactions: $\beta_j = 0$ for all j.

(IV) Truth only has main effects: $\Theta_{jk} = 0$ for all jk.

In cases (I), (II), (IV), we set 10 elements of β to be nonzero (with random sign) and in cases (I), (II), (III), we set 20 elements of the submatrix of $\Theta = \Theta^T$ to be nonzero. The signal-to-noise ratio (SNR) for the main effects part of the signal is about 1.5 whereas the SNR for the interactions part is about 1.

We study the effectiveness of the hierarchy constraint in the context of both the Lasso and forward stepwise regression. Forward stepwise regression refers to a greedy strategy for generating a sequence of linear regression models in which we start with an intercept-only model and then at each step add the variable that leads to the greatest decrease in the residual sum of squares. We choose forward stepwise as a basis of comparison since it has a simple modification that we think may be the hierarchical interactions approach most commonly used by statisticians. The modification is to restrict the set of interactions that

could be added at a given step to only those between main effect variables currently in the model. A backward stepwise version of this approach is suggested in Peixoto (1987).

We compare six methods, corresponding to each cell of the following table.

	Hierarchical	**All-Pairs**	**Main Effects Only**
Lasso	HL (our method)	APL	MEL
Fwd Stepwise	HF	APF	MEF

Each method has a single tuning parameter: for the Lasso methods, the penalty parameter, λ, and for the forward stepwise methods, the number of variables, k. We fit each method along a grid of tuning parameter values and select the model with the smallest mean square error, $E\|\hat{y}-\mu\|^2$. Note that such an operation is only possible in simulation since it requires knowing μ; however, doing so avoids the added variance of cross validation without being biased in favor of any particular method. The results presented are based on 100 simulations from the underlying model. Figure 3 shows the expected prediction error, $\sigma^2 + E[(\hat{y}-\mu)^2]$. Panel (I) shows that when the truth is hierarchical, methods that assume hierarchy (HL, HF) do better than the rest. These methods have "concentrated" their power on the correct set of models and therefore receive the biggest payoff for being correct. APL does better than MEL and MEF since it succeeds in incorporating some of the correct interactions (recall that interactions make up one quarter of the signal). In Panel (II), we notice our first surprise—that HL predicts well relative to the others even when the truth is not hierarchical! We would have expected APL (or APF) to be the clear winner in this situation since surely the hierarchy assumption can only be detrimental in this "anti-hierarchical" scenario. The reason APL doesn't outperform HL in this scenario is because APL has trouble identifying the main effects (they get swamped by the 435 interaction variables). In light of Section 3.1, this is where the hierarchy constraint helps—main effects are penalized by less and interactions by more. Even though APL is better able to find the correct interactions than HL, as seen in Panel (II) of Figure 4, APL does not predict as well as HL because it fails to find the main effects, which constitute three quarters of the signal. Relatedly, in a "hierarchical truth" scenario similar to (I) but with $p > n$ (not presented here) we have in fact observed MEL doing better than APL (though not as well as HL) since APL is not able to detect interactions accurately enough to make up for its inferior ability to detect main effects. By contrast, HL does best in that scenario, aided by hierarchy to capture both the main effect and interaction components of the signal.

In Panel (III), we see a situation where APL does dominate HL. Since there are no main effects in the signal, all that is relevant is a method's ability to find the interactions. HL identifies fewer correct interactions than APL since any main effect "information" that HL is using is spurious. Finally in Panel (IV), we see a situation where MEF, HF, MEL do better than the rest. Here again we find that the hierarchy methods beat the all-pairs methods since they favor main effects.

It is particularly illuminating to note the difference in performance between HL and HF. HF in scenarios (II), (III), and (IV) performs very similarly to the main effect only models. In (II) and (III), HL does much better than HF both in terms of prediction error and in ability to correctly identify interactions. HL appears to be far less sensitive to violations of hierarchy than HF. This difference is attributable to the joint nature in which HL acts: the decision to include a main effect is made at the same time as decisions about interactions. This allows a strong interaction to "pull" itself into the model. By contrast, HF selects main effects with no regard to the information contained in the interactions.

5.2 Data examples.

Rhee et al. (2006) study six nucleoside reverse transcriptase inhibitors (NRTIs) that are used to treat HIV-1. The target of these drugs can become resistant through mutation, and they compare a collection of models for predicting these drug's (log) susceptibility—a measure of drug resistance—based on the location of mutations. In the six cases, there are between $p = 211$ and $p = 218$ sites with mutations occurring in the $n = 784$ to $n = 1073$ samples. While they focus on main effect only models, we consider here the all-pairs lasso (APL) and weak hierarchical lasso (HL) in addition to the standard main effects lasso (MEL). We train on half of the samples and test on the remaining samples. To reduce the dependence of the results on the particular random training-test split, we repeat this process twenty times and average the results. Figure 5 shows the average test RMSE versus the average practical sparsity for each of the six drugs. In all cases but

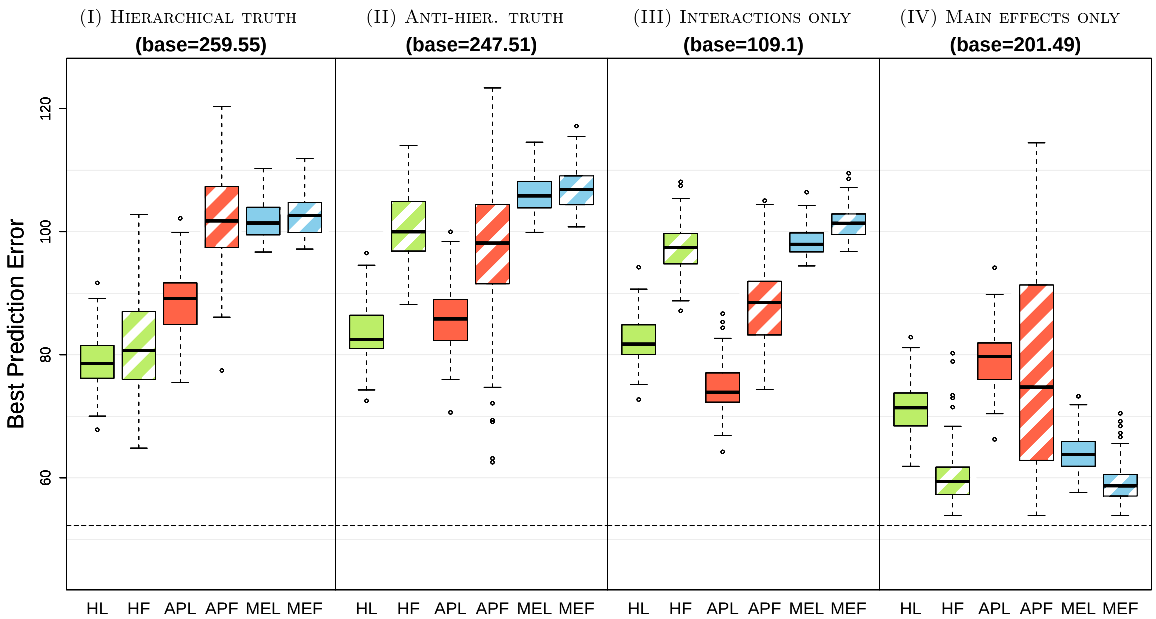

Figure 3: Prediction error: Dashed line shows Bayes error (i.e. σ^2) and the base rate refers to the prediction error of $\bar{y}_{\text{train}}$. Green, red, and blue colors indicate hierarchy, all-pairs, and main effect only, respectively; solid and striped indicate Lasso and forward stepwise, respectively.

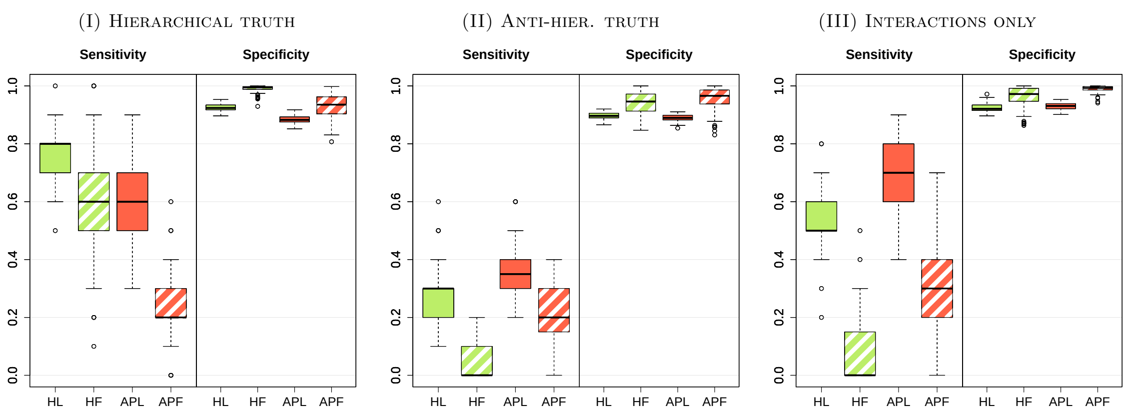

Figure 4: Plots show the ability of various methods to correctly recover the nonzero interactions. This is the sensitivity (i.e., proportion of $\Theta_{jk} \neq 0$ for which $\widehat{\Theta}_{jk} \neq 0$) and specificity (i.e., proportion of $\Theta_{jk} = 0$ for which $\widehat{\Theta}_{jk} = 0$) corresponding to the lowest prediction error model of each method.

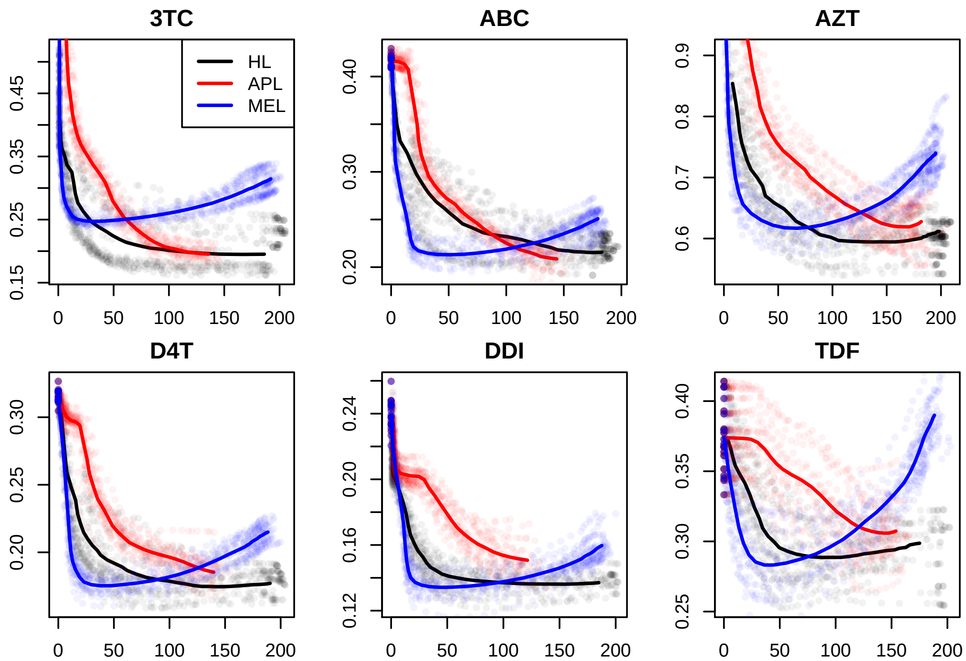

Figure 5: HIV drug data: Test-set RMSE versus practical sparsity (i.e., number of measured variables required for prediction) for six different drugs. For each method, the data from all 20 runs are displayed in faint colors; the thick lines are averages over these runs.

ABC, we find that HL achieves a better test error at most levels of practical sparsity than APL. That said, if the number of mutations one has to measure is not of concern (so that we can choose for each method the minimum RMSE model), then no method dominates in all the situations. It is worth conceding—since this is a paper on interactions—that in several of the cases a pure main effects model appears to be the best option.

6 Algorithmics.

Some of the fastest Lasso solvers rely on coordinate descent, which amounts to iteratively applying (8) until convergence (Friedman et al. 2010). Tseng (2001) proves that blockwise coordinate descent converges to the global minimum for a convex problem specifically when the nondifferentiable part of the problem is blockwise separable. In the case of the Strong Hierarchical Lasso, the hierarchy constraints combined with the symmetry constraint couples all the parameters together, meaning that coordinate descent is prone to getting stuck at suboptimal points. To see this, note that $\Theta_{jk} = \Theta_{kj}$ appears in two constraints $\|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-$ and $\|\Theta_k\|_1 \le \beta_k^+ + \beta_k^-$. By contrast, the constraints in the Weak Hierarchical Lasso problem are blockwise separable so that blockwise coordinate descent on blocks of the form $(\Theta_j, \beta_j^+, \beta_j^-)$ for $j = 1, \ldots, p$ does work. We begin by discussing our approach to solving the Weak Hierarchical Lasso problem. In Section 6.2 we discuss how we can solve a sequence of Weak Hierarchical Lasso problems that converges to a solution of the Strong Hierarchical Lasso.

6.1 Solving the Weak Hierarchical Lasso.

While blockwise coordinate descent would work for solving the Weak Hierarchical Lasso problem, we instead describe a generalized gradient descent approach. Given a problem of the form

$$\underset{\phi}{\text{Minimize}} \quad g(\phi) + h(\phi), \tag{10}$$

Algorithm 1 `WEAK-HIERNET`: Generalized gradient descent to solve Weak Hierarchical Lasso, (7), with elastic net penalty ϵ.

Inputs: $X \in \mathbb{R}^{n\times p},\ Z \in \mathbb{R}^{n\times p(p-1)},\ \lambda > 0$. Initialize $(\hat\beta^{+(0)}, \hat\beta^{-(0)}, \widehat\Theta^{(0)})$.
For $k = 1, 2, \ldots$ until convergence:
Compute residual: $\hat r^{(k-1)} \leftarrow y - X(\hat\beta^{+(k-1)} - \hat\beta^{-(k-1)}) - Z\widehat\Theta^{(k-1)}/2$.
For $j = 1, \ldots, p$:

$$\begin{aligned}(\hat\beta_j^{+(k)}, \hat\beta_j^{-(k)}, \widehat\Theta_j^{(k)}) \leftarrow \texttt{ONEROW}(&\delta\hat\beta_j^{+(k-1)} - tX_j^T\hat r^{(k-1)},\\ &\delta\hat\beta^{-(k-1)} + tX_j^T\hat r^{(k-1)},\\ &\delta\widehat\Theta_j^{(k-1)} - tZ_{(j,\cdot)}^T\hat r^{(k-1)})\end{aligned}$$

where `ONEROW` is given in Algorithm 3, $\delta = 1 - t\epsilon$, and $Z_{(j,\cdot)} \in \mathbb{R}^{n\times(p-1)}$ denotes the columns of Z involving X_j.

in which g is convex and differentiable with a Lipschitz gradient and h is convex, generalized gradient descent works by solving a sequence of problems of the form

$$\hat\phi_k \leftarrow \arg\min_\phi \frac{1}{2t}\|\phi - [\hat\phi_{k-1} - t\nabla g(\hat\phi_{k-1})]\|^2 + h(\phi),$$

where t is a suitably chosen step size (Beck & Teboulle 2009). These subproblems are easier to solve than (10) since they replace g by a spherical quadratic. Under the previously stated conditions, generalized gradient descent is guaranteed to get within $O(1/k)$ of the optimal value after k steps; in fact, with a simple modification to the algorithm, this rate improves to $O(1/k^2)$ (Beck & Teboulle 2009). Looking back at (7), we take g to be the differentiable part, $q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-)$, and h to be the ℓ_1 penalty on Θ and the set of constraints. The subproblem is of the form

$$\begin{aligned}\underset{\beta^\pm\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}}\quad & \frac{1}{2t}\|\beta^+ - \tilde\beta^+\|^2 + \frac{1}{2t}\|\beta^- - \tilde\beta^-\|^2 + \frac{1}{2t}\|\Theta - \widetilde\Theta\|_F^2 + \frac{\lambda}{2}\|\Theta\|_1\\ \text{s.t.}\quad & \left.\begin{aligned}\|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-\\ \beta_j^+ \ge 0, \beta_j^- \ge 0\end{aligned}\right\}\text{ for } j = 1, \ldots, p,\end{aligned}$$

where $(\tilde\beta^+, \tilde\beta^-, \widetilde\Theta)$ depends on the previous iteration's solution and the data, X, Z, and y. The exact form of $(\tilde\beta^+, \tilde\beta^-, \widetilde\Theta)$ is given in Algorithm 1 for solving (7), where $q(\beta_0, \beta, \Theta)$ includes an elastic net penalty as described in Remark 2 of Section 2. The above problem decouples into p separate pieces involving $(\Theta_j, \beta_j^+, \beta_j^-)$ that could be solved in parallel:

$$\begin{aligned}\underset{\beta_j^\pm\in\mathbb{R},\ \Theta_j\in\mathbb{R}^{p-1}}{\text{Minimize}}\quad & \frac{1}{2t}(\beta_j^+ - \tilde\beta_j^+)^2 + \frac{1}{2t}(\beta_j^- - \tilde\beta_j^-)^2 + \frac{1}{2t}\|\Theta_j - \widetilde\Theta_j\|^2 + \frac{\lambda}{2}\|\Theta_j\|_1\\ & \text{s.t. } \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-,\ \beta_j^+ \ge 0,\ \beta_j^- \ge 0.\end{aligned}\tag{11}$$

In Appendix E, we derive an algorithm, `ONEROW`, that solves (11) based on the observation that in terms of an optimal dual variable, $\hat\alpha$, a solution is simply $\widehat\Theta_j = \mathcal{S}[\widetilde\Theta_j,\ t(\lambda/2 + \hat\alpha)]$ and $\hat\beta_j^\pm = [\tilde\beta_j^\pm + t\hat\alpha]_+$.

We solve (7) along a sequence of λ values, from large to small, using the solution from the previous λ as a warm start for the next. The `WEAK-HIERNET` algorithm gets within ϵ of the optimal value of (7) in $O(p^2\max\{n, p\}/\epsilon)$ time.

6.2 Solving the Strong Hierarchical Lasso.

In Section 6.1, we noted that each step of generalized gradient descent conveniently decouples into p single-variable optimization problems. However, for the Strong Hierarchical Lasso, (5), the symmetry

Algorithm 2 STRONG-HIERNET: Solve (5) via ADMM.

Inputs: $X \in \mathbb{R}^{n\times p},\ Z \in \mathbb{R}^{n\times p(p-1)},\ \lambda > 0,\ \rho > 0.$

Initialize $(\hat\beta^+, \hat\beta^-, \widehat\Theta),\ \widehat\Omega,\ \widehat U.$

Repeat until convergence:

1. WEAK-HIERNET(X, Z, λ), but in the call to ONEROW replace the argument $\delta\widehat\Theta_j^{(k-1)} - tZ_{(j,\cdot)}^T \hat r^{(k-1)}$ with $\delta\widehat\Theta_j^{(k-1)} - tZ_{(j,\cdot)}^T \hat r^{(k-1)} + \rho[\widehat\Theta_j^{(k-1)} - \widehat\Omega_j] + \widehat U_j$. Also, initialize with $(\hat\beta^+, \hat\beta^-, \widehat\Theta)$.
2. $\widehat\Omega \leftarrow \frac{1}{2}(\widehat\Theta + \widehat\Theta^T) + \frac{1}{2\rho}(\widehat U + \widehat U^T)$.
3. $\widehat U \leftarrow \widehat U + \rho(\widehat\Theta - \widehat\Omega)$.

constraint ties all variables together. We therefore make use of Alternating Direction Method of Multipliers (ADMM), which is a very widely applicable framework that allows convex problems to be split apart into separate easier subproblems (Boyd et al. 2011).

Given a convex problem of the form $\text{Minimize}_\phi\ f(\phi) + g(\phi)$, we rewrite it equivalently as $\text{Minimize}_{\phi,\varphi}\ f(\phi) + g(\varphi)$ s.t. $\phi = \varphi$, and then the ADMM algorithm repeats the following three steps until convergence:

1. $\hat\phi = \text{argmin}_\phi \left[f(\phi) + (\rho/2)\|\phi - \hat\varphi + \hat u/\rho\|^2 \right]$.
2. $\hat\varphi = \text{argmin}_\varphi \left[g(\varphi) + (\rho/2)\|\varphi - \hat\phi - \hat u/\rho\|^2 \right]$.
3. $\hat u \leftarrow \hat u + \rho(\hat\phi - \hat\varphi)$.

Thus, the ADMM algorithm separates the two difficult parts of the problem, f and g, into separate optimization problems. The dual variable $\hat u$ serves to pull these two problems together, resulting in an algorithm that is guaranteed to converge to a solution as long as $\rho > 0$. In practice, the value of ρ affects the speed of convergence.

In our case, we use ADMM to separate the hierarchy constraints, involving $(\beta^+, \beta^-, \Theta)$ from the symmetry constraint, which will involve a symmetric version of Θ, which we call Ω:

$$\begin{aligned} \underset{\substack{\beta_0\in\mathbb{R},\ \beta^\pm\in\mathbb{R}^p,\\ \Theta,\Omega\in\mathbb{R}^{p\times p}}}{\text{Minimize}} \quad & q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad \Omega = \Omega^T, \Theta = \Omega \quad & \left.\begin{aligned} \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^- \\ \beta_j^+ \ge 0, \beta_j^- \ge 0 \end{aligned}\right\} \text{for } j = 1, \dots, p. \end{aligned} \tag{12}$$

The resulting ADMM algorithm is given in Algorithm 2, which is explained in greater detail in Appendix G. Conceptually, the algorithm alternately updates two matrices, Θ and Ω. Throughout the algorithm, we update Θ by solving a version of problem (7), and we update Ω by symmetrizing a version of Θ. At convergence, $\widehat\Theta = \widehat\Omega$, and thus $\widehat\Theta$ is both symmetric and satisfies the hierarchy constraints.

7 Discussion.

In this paper, we have proposed a modification to the Lasso for fitting strong and weak hierarchical interaction models. These two approaches are closely tied and our algorithms to solve the two exploit their similar structure. A key advantage of our framework is that it admits a simple characterization of the effect of imposing hierarchy. We compare our hierarchical methods to the Lasso and stepwise procedures to understand the implications of demanding hierachy. We introduce a distinction between models that have a small number of parameters and those that require measuring only a small number of variables. The

hierarchical interaction requirement favors models with the latter type of sparsity, a feature that is desirable when performing measurements is costly, time consuming, or otherwise inconvenient. The `R` package `hierNet` provides implementations of our strong and weak methods, both for Gaussian and logistic losses. This work has potential applications to genomewide association studies. In future work, we intend to extend this framework to contexts in which only certain interactions should be considered such as in gene-environment interaction models.

8 Acknowledgments.

We thank Will Fithian, Max Grazier-G'Sell, Brad Klingenberg, Balasubramanian Narasimhan, and Ryan Tibshirani for useful conversations and two referees and an associate editor for helpful comments. Jacob Bien was supported in part by the Gerald J. Lieberman Fellowship, Jonathan Taylor by National Science Foundation Grant DMS-0906801 and AFOSR Grant 113039, and Robert Tibshirani by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

Appendix

A Proof that (5) and (6) are equivalent.

We rewrite (5) in terms of $\beta = \beta^+ - \beta^-$ rather than β^-:

$$\begin{aligned} \underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta) + \lambda 1^T(2\beta^+ - \beta) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad & \Theta = \Theta^T,\ \beta^+ \geq 0,\ \beta^+ \geq \beta,\ \|\Theta_j\|_1 \leq 2\beta_j^+ - \beta_j. \end{aligned}$$

or

$$\begin{aligned} \underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta) + \lambda 1^T(2\beta^+ - \beta) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad & \Theta = \Theta^T,\ \max\{[\beta_j]_+,\ (\|\Theta_j\|_1 + \beta_j)/2\} \leq \beta_j^+, \end{aligned}$$

where $[\beta_j]_+ = \max\{\beta_j, 0\}$ is the positive part of β_j. This problem is the epigraph form of

$$\begin{aligned} \underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta) + \lambda\sum_{j=1}^p(\max\{2[\beta_j]_+, \|\Theta_j\|_1 + \beta_j\} - \beta_j) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad & \Theta = \Theta^T \end{aligned}$$

which reduces to (6) since $2[\beta_j]_+ - \beta_j = |\beta_j|$.

B Effect of constraint.

For notational simplicity, we write $r(\beta^+,\beta^-,\Theta) \in \mathbb{R}^n$ to denote the residuals, $y - \hat{y}(\beta^+,\beta^-,\Theta)$, as a function of the parameters. The strong Hierarchical Lasso problem is the following:

$$\begin{aligned} \text{minimize}_{\beta^+,\beta^-,\Theta} \quad & \frac{1}{2}\|r(\beta^+,\beta^-,\Theta)\|^2 + \lambda_1 1^T(\beta^+ + \beta^-) + \lambda_2\sum_j\|\Theta_j\|_1 \\ \text{s.t.} \quad & \|\Theta_j\|_1 \leq \beta_j^+ + \beta_j^- \text{ and } \beta_j^+ \geq 0,\ \beta_j^- \geq 0 \text{ for each } j,\ \Theta = \Theta^T \end{aligned}$$

The Lagrangian is

$$\begin{aligned} L(\phi;\alpha,S,\gamma^\pm,U) = & \frac{1}{2}\|r(\beta^+,\beta^-,\Theta)\|^2 + \lambda_1 1^T(\beta^+ + \beta^-) + \lambda_2\langle U,\Theta\rangle + \sum_j\alpha_j(U_j^T\Theta_j - \beta_j^+ - \beta_j^-) \\ & - \gamma_j^+\beta_j^+ - \gamma_j^-\beta_j^- + \langle S, \Theta - \Theta^T\rangle \\ = & \frac{1}{2}\|r(\beta^+,\beta^-,\Theta)\|^2 + (\lambda_1 1 - \alpha - \gamma^+)^T\beta^+ + (\lambda_1 1 - \alpha - \gamma^-)^T\beta^- \\ & + \langle S - S^T + \operatorname{diag}(\lambda_2 1 + \alpha)U, \Theta\rangle. \end{aligned}$$

where $\alpha, \gamma^{\pm}, S, U$ are dual variables. According to the KKT conditions, $(\hat{\phi}; \hat{\alpha}, \widehat{S}, \hat{\gamma}^{\pm}, \widehat{U})$ is an optimal primal-dual pair if and only if

$$\begin{aligned} \pm x_j^T r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta}) &= \lambda_1 - \hat{\alpha}_j - \hat{\gamma}_j^{\pm} \\ (x_j * x_k)^T r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})/2 &= (\lambda_2 + \hat{\alpha}_j)\widehat{U}_{jk} + \widehat{S}_{jk} - \widehat{S}_{kj} \\ 0 = \hat{\beta}_j^{\pm}\hat{\gamma}_j^{\pm} & \qquad 0 = \hat{\alpha}_j(\|\widehat{\Theta}_j\|_1 - \hat{\beta}_j^+ - \hat{\beta}_j^-) \\ \widehat{\Theta} = \widehat{\Theta}^T, & \qquad \hat{\beta}^{\pm} \geq 0, \qquad \|\widehat{\Theta}_j\|_1 \leq \hat{\beta}_j^+ + \hat{\beta}_j^- \qquad \hat{\alpha}, \hat{\gamma}^{\pm} \geq 0 \end{aligned}$$

$$\widehat{U}_{jk} = \begin{cases} \text{sign}(\widehat{\Theta}_{jk}) & \widehat{\Theta}_{jk} \neq 0 \\ \in [-1, 1] & \widehat{\Theta}_{jk} = 0. \end{cases}$$

Now, letting $r^{(-j)} = r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta}) + (\hat{\beta}_j^+ - \hat{\beta}_j^-)x_j$ and recalling that $\|x_j\|^2 = 1$, there are three cases to consider:

1. $\hat{\beta}_j^+ \geq 0,\ \hat{\beta}_j^- = 0$:

$$x_j^T(r^{(-j)} - \hat{\beta}_j^+ x_j) = \lambda_1 - \hat{\alpha}_j - \hat{\gamma}_j^+ \implies \hat{\beta}_j^+ = [x_j^T r^{(-j)} - (\lambda_1 - \hat{\alpha}_j)]_+$$

   Note that this case applies when $x_j^T r^{(-j)} \geq \lambda_1 - \hat{\alpha}_j$. Thus, in this case $\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat{\alpha}_j)$.

2. $\hat{\beta}_j^+ = 0,\ \hat{\beta}_j^- \geq 0$:

$$-x_j^T(r^{(-j)} + \hat{\beta}_j^- x_j) = \lambda_1 - \hat{\alpha}_j - \hat{\gamma}_j^- \implies \hat{\beta}_j^- = [-x_j^T r^{(-j)} - (\lambda_1 - \hat{\alpha}_j)]_+$$

   Note that this case applies when $x_j^T r^{(-j)} \leq -(\lambda_1 - \hat{\alpha}_j)$. Thus, once again $\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat{\alpha}_j)$.

3. $\hat{\beta}_j^+ > 0,\ \hat{\beta}_j^- > 0 \quad (\implies \hat{\gamma}_j^+ = 0,\ \hat{\gamma}_j^- = 0)$

$$\pm x_j^T(r^{(-j)} - (\hat{\beta}_j^+ - \hat{\beta}_j^-)x_j) = \lambda_1 - \hat{\alpha}_j \implies \hat{\beta}_j^+ - \hat{\beta}_j^- = x_j^T r^{(-j)}$$

   Note that this case applies when $\hat{\alpha}_j = \lambda_1$, so trivially $\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat{\alpha}_j)$.

Thus, we have shown that $\hat{\beta}_j^+ - \hat{\beta}_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat{\alpha}_j)$.

We can get rid of $\hat{S}$ by rewriting the subgradient equation involving it as

$$(x_j * x_k)^T r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta}) = (2\lambda_2 + \hat{\alpha}_j + \hat{\alpha}_k)\widehat{U}_{jk}$$

(note that symmetry in $\widehat{\Theta}$ implies that there exists a symmetric $\widehat{U}$).

Now, letting $r^{(-jk)} = r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta}) + (x_j * x_k)(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2$, we get

$$\widehat{\Theta}_{jk}\|x_j * x_k\|^2 = (x_j * x_k)^T r^{(-jk)} - (2\lambda_2 + \hat{\alpha}_j + \hat{\alpha}_k)\widehat{U}_{jk} = \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2\lambda_2 + \hat{\alpha}_j + \hat{\alpha}_k).$$

This completes the proof for the Strong Hierarchical Lasso. Note that in the Weak Hierarchical Lasso case, the KKT conditions are identical except we do not have the constraint $\widehat{\Theta} = \widehat{\Theta}^T$ and we take $\widehat{S} = 0$. Thus, the relevant condition is simply

$$(x_j * x_k)^T r(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta}) = 2(\lambda_2 + \hat{\alpha}_j)\widehat{U}_{jk} = 2(\lambda_2 + \hat{\alpha}_k)\widehat{U}_{kj}.$$

Note that the second equality implies that $\widehat{U}_{jk}\widehat{U}_{kj} \geq 0$ (since $\hat{\alpha} \geq 0$) and that if $|U_{jk}| = 1$, then $\hat{\alpha}_j \leq \hat{\alpha}_k$ and vice versa. Rearranging terms, we have

$$\begin{aligned} (\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j)\widehat{U}_{jk} \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_k)\widehat{U}_{kj} \end{aligned}$$

Now, $\widehat{U}_{jk}\widehat{U}_{kj} \geq 0$ implies $\widehat{\Theta}_{jk}\widehat{\Theta}_{kj} \geq 0$ which implies that $(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2$, if nonzero, has the same sign as whichever of $\widehat{\Theta}_{jk}$ or $\widehat{\Theta}_{kj}$ (or both) is nonzero.

There are four cases:

1. $\widehat{\Theta}_{jk} \neq 0$, $\widehat{\Theta}_{kj} = 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_j \leq \hat{\alpha}_k$ since $|\widehat{U}_{jk}| = 1$.

2. $\widehat{\Theta}_{jk} = 0$, $\widehat{\Theta}_{kj} \neq 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_k) \cdot \mathrm{sign}(\widehat{\Theta}_{kj}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_k) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_k \leq \hat{\alpha}_j$ since $|\widehat{U}_{kj}| = 1$.

3. $\widehat{\Theta}_{jk} \neq 0$, $\widehat{\Theta}_{kj} \neq 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_j = \hat{\alpha}_k$ since $|\widehat{U}_{jk}| = |\widehat{U}_{kj}| = 1$.

4. $\widehat{\Theta}_{jk} = 0$, $\widehat{\Theta}_{kj} = 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= 0 \\ &= \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \hat{\alpha}_j)) \\ &= \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \hat{\alpha}_k))\end{aligned}$$
where the latter two equalities follow since $|(x_j * x_k)^T r^{(-jk)}| \leq 2(\lambda_2 + \hat{\alpha}_j)$ and $|(x_j * x_k)^T r^{(-jk)}| \leq 2(\lambda_2 + \hat{\alpha}_k)$

We can encapsulate all of this into a single, simple expression:
$$(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 = \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \min\{\hat{\alpha}_j, \hat{\alpha}_k\}))$$

C Proofs of strong and weak hierarchy.

We begin by proving Lemma 1, which characterizes all solutions to (5) as a relatively simple function of y. The structure of our proof is based on (Ryan) Tibshirani & Taylor (2011, 2012).

C.1 Characterizing the solution.

For ease of analysis, we write (5) equivalently in terms of Θ^+ and Θ^-. Also, for notational simplicity, we write $\phi = (\beta^+, \beta^-, \Theta^+, \Theta^-)$ and $\widetilde{X} = (X; -X; Z/2; -Z/2)$. The strong Hierarchical Lasso problem is the following:
$$\begin{aligned}\mathrm{minimize}_{\beta^+,\beta^-,\Theta^+,\Theta^-} \quad & \frac{1}{2}\|y - \widetilde{X}\phi\|^2 + \lambda_1 1^T(\beta^+ + \beta^-) + \lambda_2\langle 11^T, \Theta^+ + \Theta^-\rangle \\ \text{s.t.} \quad & 1^T(\Theta_j^+ + \Theta_j^-) \leq \beta_j^+ + \beta_j^- \text{ and } \beta_j^+ \geq 0,\ \beta_j^- \geq 0 \text{ for each } j, \\ & \Theta^+ - \Theta^- = \Theta^{+T} - \Theta^{-T},\ \Theta_{jk}^\pm \geq 0,\ \Theta_{jj}^\pm = 0\end{aligned}$$

In introducing $\Theta^{\pm}$, we are not in fact changing the problem since at a solution $\widehat{\Theta}^{\pm} = \max\{\pm\widehat{\Theta}, 0\}$ (for $\lambda_2 > 0$). To see this, note that given any feasible point with $\Theta^+_{jk} > 0$ and $\Theta^-_{jk} > 0$, we can produce a feasible point with strictly lower objective by reducing $\Theta^+_{jk}, \Theta^-_{jk}, \Theta^+_{kj}, \Theta^-_{kj}$ all by equal amounts.

We will try to make this as close in form and notation to (Ryan) Tibshirani & Taylor (2012)'s treatment of the generalized lasso problem as possible. Our optimization problem is of the form

$$\text{minimize}_{\phi} \quad \frac{1}{2}\|y - \widetilde{X}\phi\|^2 + w^T\phi \quad \text{s.t.} \quad D\phi \geq 0, \quad L\phi = 0. \tag{13}$$

The Lagrangian of this problem is

$$L(\phi; \mu, \nu) = \frac{1}{2}\|y - \widetilde{X}\phi\|^2 + w^T\phi - \mu^T D\phi + \nu^T L\phi,$$

where $\mu \geq 0$ and ν are dual variables. The KKT conditions for $(\hat{\phi}(y), (\hat{\mu}(y), \hat{\nu}(y)))$ to be an optimal primal-dual pair are the following:

$$\begin{aligned}
\widetilde{X}^T(y - \widetilde{X}\hat{\phi}) &= w - D^T\hat{\mu} + L^T\hat{\nu} \\
\hat{\mu}_i(D\hat{\phi})_i &= 0 \\
\hat{\mu} &\geq 0 \\
D\hat{\phi} \geq 0, L\hat{\phi} &= 0.
\end{aligned}$$

Now, define the "boundary" and "active" sets as

$$\begin{aligned}
\mathcal{B}(\hat{\mu}) &= \{i : \hat{\mu}_i = 0\} \\
\mathcal{A}(\hat{\phi}) &= \{i : [D\hat{\phi}]_i > 0\}.
\end{aligned}$$

These are not necessarily unique since $(\hat{\phi}, (\hat{\mu}, \hat{\nu}))$ may not be unique. In terms of the active set $\mathcal{A}(\hat{\phi})$, the KKT conditions become

$$\begin{aligned}
\widetilde{X}^T(y - \widetilde{X}\hat{\phi}) &= w - D^T_{-\mathcal{A}(\hat{\phi})}\hat{\mu}_{-\mathcal{A}(\hat{\phi})} + L^T\hat{\nu} \\
\hat{\mu}_{\mathcal{A}(\hat{\phi})} &= 0 \qquad \hat{\mu} \geq 0 \\
L\hat{\phi} &= 0 \qquad D_{-\mathcal{A}(\hat{\phi})}\hat{\phi} = 0.
\end{aligned}$$

Solving for $\hat{\phi}$, we get the following characterization of a Strong Hierarchical Lasso solution:

Lemma 1. *Suppose $\hat{\phi}$ is a solution to the Strong Hierarchical Lasso problem* (5) *(taking $q(\beta_0, \beta, \Theta)$ to be the quadratic loss) with $\mathcal{A}(\hat{\phi}) = \{i : [D\hat{\phi}]_i > 0\}$. Then, $\hat{\phi}$ can be written in terms of $\mathcal{A}(\hat{\phi})$ and y as*

$$\hat{\phi} = (\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}^T)^+w) + b,$$

where $b \in \text{null}(\widetilde{X}) \cap \text{null}(L) \cap \text{null}(D_{-\mathcal{A}(\hat{\phi})})$ satisfies

$$D_i[(\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}^T)^+w) + b] > 0$$

for all $i \in \mathcal{A}(\hat{\phi})$.

Proof. Defining $\widetilde{D} = \begin{pmatrix} D_{-\mathcal{A}(\hat{\phi})} \\ L \end{pmatrix}$ and $P = P_{\text{null}(\widetilde{D})} = P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}$, we solve for $\hat{\phi}$ in the same manner as is done in (Ryan) Tibshirani & Taylor (2012). Since $\widetilde{D}\hat{\phi} = 0$ is equivalent to $P\hat{\phi} = \hat{\phi}$, we have $P\widetilde{X}^T(y -$

$\widetilde{X}P\hat{\phi}) = Pw$. We see that $Pw \in \text{col}(P\widetilde{X}^T)$ and thus $Pw = (P\widetilde{X}^T)(P\widetilde{X}^T)^+Pw$. Thus, $P\widetilde{X}^T\widetilde{X}P\hat{\phi} = P\widetilde{X}^T(y - (P\widetilde{X}^T)^+Pw)$ from which we get

$$\hat{\phi} = (\widetilde{X}P)^+(y - (P\widetilde{X}^T)^+Pw) + b,$$

for $b \in \text{null}(\widetilde{X}P)$ and such that $\widetilde{D}b = 0$ and $D_i\hat{\phi} > 0$ for $i \in \mathcal{A}(\hat{\phi})$. To complete the result, we observe that the first two conditions reduce to $b \in \text{null}(\widetilde{X}) \cap \text{null}(L) \cap \text{null}(D_{-\mathcal{A}(\hat{\phi})})$. □

We will use this characterization of a solution both to prove that the hierarchy property holds with probability one under weak assumptions and to derive an unbiased estimate of the degrees of freedom.

Before we do so, we write out $\widetilde{D} = \begin{pmatrix} D_{-\mathcal{A}(\hat{\phi})} \\ L \end{pmatrix}$ more explicitly and introduce a little notation that will be useful later. Every row of D corresponds to an inequality constraint, and we can partition these rows into ten subsets:

$$\begin{aligned} \mathcal{L} &= \{j : 1^T(\widehat{\Theta}_j^+ + \widehat{\Theta}_j^-) < \hat{\beta}_j^+ + \hat{\beta}_j^-\} & \mathcal{T} &= \mathcal{L}^c \\ \mathcal{P}(\hat{\beta}^\pm) &= \{j : \hat{\beta}_j^\pm > 0\} & \mathcal{Z}(\hat{\beta}^\pm) &= \mathcal{P}(\hat{\beta}^\pm)^c \\ \mathcal{P}(\widehat{\Theta}^\pm) &= \{j \neq k : \widehat{\Theta}_{jk}^\pm > 0\} & \mathcal{Z}(\widehat{\Theta}^\pm) &= \mathcal{P}(\widehat{\Theta}^\pm)^c \end{aligned} \tag{14}$$

The set $\mathcal{A}(\hat{\phi})^c$ is made up of $\mathcal{T}$, $\mathcal{Z}(\hat{\beta}^+)$, $\mathcal{Z}(\hat{\beta}^-)$, $\mathcal{Z}(\widehat{\Theta}^+)$, and $\mathcal{Z}(\widehat{\Theta}^-)$. The matrix $\widetilde{D}$ has $2p + 2p^2$ columns that can be partitioned as $(\widetilde{D}^{\beta^+} : \widetilde{D}^{\beta^-} : \widetilde{D}^{\Theta^+} : \widetilde{D}^{\Theta^-})$ and a row for every constraint. The rows of this matrix are the following (where e_j and 1_p are row vectors):

R1. For each $j \in \mathcal{T}$,	(e_j	e_j	$-e_j \otimes 1_p$	$-e_j \otimes 1_p$	)
R2. For each $j \in \mathcal{Z}(\beta^+)$,	(e_j	0	0	0	)
R3. For each $j \in \mathcal{Z}(\beta^-)$,	(0	e_j	0	0	)
R4. For each $jk \in \mathcal{Z}(\Theta^+)$,	(0	0	$e_j \otimes e_k$	0	)
R5. For each $jk \in \mathcal{Z}(\Theta^-)$,	(0	0	0	$e_j \otimes e_k$	)
R6. For each j,	(0	0	$e_j \otimes e_j$	0	)
R7. For each j,	(0	0	0	$e_j \otimes e_j$	)
R8. For each $j < k$,	(0	0	$e_j \otimes e_k + e_k \otimes e_j$	$-e_j \otimes e_k - e_k \otimes e_j$	)

We will refer to this in the proofs that follow.

C.2 Proof of strong hierarchy (Property 2).

Including the elastic net penalty, $(\epsilon/2)\|\beta^+\|^2 + (\epsilon/2)\|\beta^-\|^2 + (\epsilon/2)\|\Theta\|_F^2$, is equivalent to replacing $\widetilde{X}$ and y in (13) by

$$\widetilde{X}_\epsilon = \begin{pmatrix} X & -X & Z/2 & -Z/2 \\ \sqrt{\epsilon}I_p & 0 & 0 & 0 \\ 0 & \sqrt{\epsilon}I_p & 0 & 0 \\ 0 & 0 & \sqrt{\epsilon}I_{p^2} & -\sqrt{\epsilon}I_{p^2} \end{pmatrix} \text{ and } y_\epsilon = \begin{pmatrix} y \\ 0_{(2p+p^2)\times 1} \end{pmatrix}.$$

Suppose we solve (13) with the above design matrix. By Lemma 1,

$$\hat{\phi} = (\widetilde{X}_\epsilon P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}_\epsilon^T)^+w) + b,$$

for some $b \in \text{null}(\widetilde{X}_\epsilon) \cap \text{null}(L) \cap \text{null}(D_{-\mathcal{A}(\hat{\phi})})$ satisfying

$$D_i[(\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}^T)^+w) + b] > 0$$

for all $i \in \mathcal{A}(\hat{\phi})$. Let $S_v : \mathbb{R}^{2p+2p^2} \to \mathbb{R}^{|v|}$ be the linear operator that selects the part of a vector corresponding to the variable v. Now, $b \in \text{null}(\widetilde{X}_\epsilon)$ implies that $S_{\beta^+}(b) = S_{\beta^-}(b) = 0$ and $S_{\Theta^+}(b) = -S_{\Theta^-}(b)$. We showed earlier that we cannot have $\widehat{\Theta}_{jk}^+ > 0$ and $\widehat{\Theta}_{jk}^- > 0$. This means that for any jk, there must be an $i \notin \mathcal{A}(\hat{\phi})$

for which $D_i\hat{\phi} = 0$ corresponds to $S_{\Theta^+_{jk}}(\hat{\phi}) = 0$ or $S_{\Theta^-_{jk}}(\hat{\phi}) = 0$. Thus, $D_{-\mathcal{A}(\hat{\phi})}b = 0$ means that $S_{\Theta^+_{jk}}(b) = 0$ or $S_{\Theta^-_{jk}}(b) = 0$ for each jk. This implies that $\text{null}(\widetilde{X}_\epsilon) \cap \text{null}(D_{-\mathcal{A}(\hat{\phi})}) = \{0\}$ and thus $b = 0$.

We show now that $P\left(\bigcup_j\{\hat{\beta}^+_j = \hat{\beta}^-_j > 0\}\right) = 0$. In terms of our above notation, this is

$$P\left(\bigcup_j\{\hat{\phi}^T(e^+_j - e^-_j) = 0,\ \hat{\phi}^Te^+_j > 0,\ \hat{\phi}^Te^-_j > 0\}\right) = 0, \tag{15}$$

where $e^\pm_j \in \mathbb{R}^{2p+2p^2}$ is the vector with all zeros except for $S_{\beta^\pm}(e^\pm_j) = 1$. Consider the set

$$\begin{aligned}\mathcal{N} = \cup_{\mathcal{A}} \cup_{j=1}^p \{z :&[(\widetilde{X}_\epsilon P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})})^+(z - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})}\widetilde{X}^T_\epsilon)^+w)]^T(e^+_j - e^-_j) = 0\\ &[(\widetilde{X}_\epsilon P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})})^+(z - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})}\widetilde{X}^T_\epsilon)^+w)]^Te^+_j > 0\\ &[(\widetilde{X}_\epsilon P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})})^+(z - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})}\widetilde{X}^T_\epsilon)^+w)]^Te^-_j > 0\},\end{aligned}$$

In light of (14), fixing $\mathcal{A}$ automatically specifies $\mathcal{T}, \mathcal{P}(\hat{\beta}^\pm), \mathcal{P}(\widehat{\Theta}^\pm)$. The outer union is restricted to those subsets $\mathcal{A}$ of $2p + 2p^2$ elements that would have $\mathcal{P}(\widehat{\Theta}^+) \cap \mathcal{P}(\widehat{\Theta}^-) = \emptyset$ and $\mathcal{P}(\hat{\beta}^+) \cap \mathcal{P}(\hat{\beta}^-) \subseteq \mathcal{T}$. The event in (15) is contained in $\{y_\epsilon \in \mathcal{N}\}$ since it corresponds to the case in which $\mathcal{A}$ is $\mathcal{A}(\hat{\phi})$. We begin by showing that $\mathcal{A}(\hat{\phi})$ is in this restricted union with probability one. We have already argued that at a solution we must have $\widehat{\Theta}^+_{jk}\widehat{\Theta}^-_{jk} = 0$ for all jk. Now, $\hat{\beta}^+_j > 0$ and $\hat{\beta}^-_j > 0$ together imply that $1^T(\widehat{\Theta}^+_j + \widehat{\Theta}^-_j) = \hat{\beta}^+_j + \hat{\beta}^-_j$ since otherwise we could lower the objective by reducing $\hat{\beta}^+_j$ and $\hat{\beta}^-_j$ without leaving the feasible set. Therefore, it would be sufficient to show that $P(y_\epsilon \in \mathcal{N}) = 0$. We do so by observing that $\{y_\epsilon \in \mathcal{N}\}$ is a finite union of zero probability sets.

We begin by establishing that $\text{null}(I_{1:n}(P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})}\widetilde{X}^T_\epsilon)^+) = \text{null}(\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})})$. To do so, we write $P = P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}})}$ as $P = UU^T$ for some $U^TU = I$. Now, $\text{row}(\widetilde{X}P) \subseteq \text{row}(P) = \text{col}(U)$, so we can write $U = (U_1 : U_2)$, where $\text{row}(\widetilde{X}P) = \text{col}(U_1)$ and thus $\widetilde{X}PU_2 = 0$. Since $PU_2 = U_2$, it follows that $\widetilde{X}U_2 = 0$. Write $U^T_i = (U^{\beta^+T}_i : U^{\beta^-T}_i : U^{\Theta^+T}_i : U^{\Theta^-T}_i)$ for $i = 1, 2$, and observe that $U^T_i\widetilde{X}^T_\epsilon = [U^T_i\widetilde{X}^T : \sqrt{\epsilon}U^{\beta^+T}_i : \sqrt{\epsilon}U^{\beta^-T}_i : \sqrt{\epsilon}(U^{\Theta^+T}_i - U^{\Theta^-T}_i)]$ so that

$$\begin{aligned}U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_2 &= 0 + \epsilon[U^{\beta^+T}_1U^{\beta^+}_2 + U^{\beta^-T}_1U^{\beta^-}_2 + (U^{\Theta^+}_1 - U^{\Theta^-}_1)(U^{\Theta^+}_2 - U^{\Theta^-}_2)]\\ &= \epsilon[U^T_1U_2 - U^{\Theta^+T}_1U^{\Theta^-}_2 - U^{\Theta^-T}_1U^{\Theta^+}_2]\\ &= -\epsilon[0 + \sum_{jk}[U^{\Theta^+}_1]_{jk}[U^{\Theta^-}_2]_{jk} + [U^{\Theta^-}_1]_{jk}[U^{\Theta^+}_2]_{jk}].\end{aligned}$$

Now, for each jk we must have $jk \in \mathcal{Z}(\Theta^+) \cup \mathcal{Z}(\Theta^-)$ since $\mathcal{P}(\Theta^+) \cap \mathcal{P}(\Theta^-) = \emptyset$. Thus, for each jk, there is a R4 or R5 row in $D_{-\mathcal{A}}$ and thus $D_{-\mathcal{A}}(U_1 : U_2) = 0$ implies that $[U^{\Theta^+}_1]_{jk} = 0$ or $[U^{\Theta^-}_2]_{jk} = 0$ for each jk and likewise $[U^{\Theta^+}_2]_{jk} = 0$ or $[U^{\Theta^-}_1]_{jk} = 0$. Therefore, $U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_2 = 0$. Now,

$$\begin{aligned}I_{1:n}(P\widetilde{X}^T_\epsilon)^+ &= \widetilde{X}P(P\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon P)^+ = \widetilde{X}UU^T(UU^T\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon UU^T)^+\\ &= \widetilde{X}UU^TU(U^T\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U)^+U^T = \widetilde{X}U(U^T\widetilde{X}^T_{\mathcal{A}}\widetilde{X}_\epsilon U)^+U^T\\ &= \widetilde{X}(U_1\ U_2)\begin{pmatrix}U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_1 & U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_2\\ U^T_2\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_1 & U^T_2\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_2\end{pmatrix}^+U^T\\ &= (\widetilde{X}U_1\ 0)\begin{pmatrix}U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_1 & 0\\ 0 & U^T_2\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_2\end{pmatrix}^+U^T\\ &= \widetilde{X}U_1(U^T_1\widetilde{X}^T_\epsilon\widetilde{X}_\epsilon U_1)^+U^T_1.\end{aligned}$$

Now, $U_1^T \widetilde{X}_\epsilon^T \widetilde{X}_\epsilon U_1 \succ 0$ since $U_1^T \widetilde{X}_\epsilon^T \widetilde{X}_\epsilon U_1 = U_1^T \widetilde{X}^T \widetilde{X} U_1 + J$ and $J \succeq 0$ and $\widetilde{X} U_1$ has full column rank. Thus, $\text{null}(I_{1:n}(P\widetilde{X}_\epsilon^T)^+) = \text{null}(U_1^T) = \text{null}(\widetilde{X}P)$. This completes the first part of the proof.

Next, we show that $\widetilde{X}P(e_j^+ - e_j-) \neq 0$ as long as $(\widetilde{X}_\epsilon P)^+(y_\epsilon - (P\widetilde{X}_\epsilon^T)^+ w)]^T e_j^\pm > 0$. Now, since $j \in \cap\mathcal{P}(\beta^+) \cap \mathcal{P}(\beta^-) \subseteq \mathcal{T}$ and so the only row of $\widetilde{D}$ that has $\widetilde{D}_i(e_j^\pm) \neq 0$ is the R1 row; but clearly $\widetilde{D}_i(e_j^+ - e_j^-) = 1 - 1 = 0$. Thus, $e_j^+ - e_j^- \in \text{null}(\widetilde{D})$ and $P(e_j^+ - e_j^-) = e_j^+ - e_j^-$. It follows that

$$\widetilde{X}P(e_j^+ - e_j^-) = \begin{pmatrix} 2x_j \\ \sqrt{\epsilon} e_j \\ -\sqrt{\epsilon} e_j \\ 0 \end{pmatrix} \neq 0 \text{ assuming } \epsilon > 0.$$

Putting these two parts of the proof together establishes that $I_{1:n}(P\widetilde{X}_\epsilon^T)^+(e_j^+ - e_j^-) \neq 0$. Thus, $\{y_\epsilon \in \mathcal{N}\}$ is a finite union of Lebesgue measure 0 sets. This shows that $P(y_\epsilon \in \mathcal{N}) = 0$ as long as y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$.

C.3 Proof of weak hierarchy (Property 3).

An argument nearly identical to that of the previous section establishes that $\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0$ with probability one. Thus, if $\hat{\beta}_j^+ - \hat{\beta}_j^- = 0$ and $\hat{\beta}_k^+ - \hat{\beta}_j^- = 0$, then both $\widehat{\Theta}_{jk} = 0$ and $\widehat{\Theta}_{kj} = 0$. It follows then that $(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2 = 0$. This establishes weak hierarchy.

D Degrees of freedom.

D.1 Proof of unbiased estimate (Property 4).

The fit in terms of the active set is given by

$$\widetilde{X}\hat{\phi} = (\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})(\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}^T)^+ w)$$

Of course, $\hat{\mu}_{\mathcal{A}(\hat{\phi})} = 0$ and we can solve the KKT conditions to get the rest of the optimal dual variables in terms of the active set:

$$\begin{pmatrix} \hat{\mu}_{-\mathcal{A}(\hat{\phi})} \\ \hat{\nu} \end{pmatrix} = \widetilde{D}^{T+}[w - \widetilde{X}^T(y - \widetilde{X}\hat{\phi})] + c$$

where $c \in \text{null}(\widetilde{D})$ satisfies $\widetilde{D}^{T+}[w - \widetilde{X}^T(y - \widetilde{X}\hat{\phi})] + c \geq 0$.

Note that $\hat{\phi} = \hat{\phi}(y)$ and thus $\mathcal{A}(\hat{\phi})$ and b depend on y even though we don't write this explicitly. We will continue writing $\hat{\phi}$ to mean specifically $\hat{\phi}(y)$. For y' in a neighborhood of y, we might guess that $\hat{\phi}(y') = f(y')$ and $(\hat{\mu}(y'), \hat{\nu}(y')) = (g(y'), h(y'))$, where

$$\begin{aligned}
f(y') &= (\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+(y' - (P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}\widetilde{X}^T)^+ w) + b \\
\begin{pmatrix} g(y')_{-\mathcal{A}(\hat{\phi})} \\ h(y') \end{pmatrix} &= \begin{pmatrix} D_{-\mathcal{A}(\hat{\phi})} \\ L \end{pmatrix}^{T+} [w - \widetilde{X}^T(y' - \widetilde{X}f(y'))] + c \\
g(y')_{\mathcal{A}(\hat{\phi})} &= 0
\end{aligned}$$

To verify this guess, we need to check that the pair $(f(y'), (g(y'), h(y')))$ satisfies the optimality conditions at y':

$$\begin{aligned}
\widetilde{X}^T(y' - \widetilde{X}f(y')) &= w - D^T g(y') + L^T h(y') \\
g_i(y')(Df(y'))_i &= 0 \\
g(y') \geq 0, \qquad Df(y') &\geq 0, \qquad Lf(y') = 0.
\end{aligned}$$

First of all, $Lf(y') = 0$ holds since $L(\widetilde{X}P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})})^+ = 0$ and $Lb = 0$. Likewise, $D_{-\mathcal{A}(\hat{\phi})}f(y') = 0$. Now, $D_{\mathcal{A}(\hat{\phi})}f(y) > 0$ so by continuity of f, we have $D_{\mathcal{A}(\hat{\phi})}f(y') > 0$ for all y' in a small enough neighborhood, U_1, of y. This establishes that $\mathcal{A}(f(y')) = \mathcal{A}(\hat{\phi})$. Now, $g(y')_{\mathcal{A}(\hat{\phi})} = 0$, so complementary slackness holds. To see that the first optimality condition holds, we can simply plug $(g(y'), h(y'))$ into the left hand side. All that remains is to show that $g(y')_{-\mathcal{A}(\hat{\phi})} \geq 0$. If we knew that $\hat{\mu}_{-\mathcal{A}(\hat{\phi})} > 0$, then by continuity of g we could argue that over a small enough neighborhood, U_2, $g(y')_{-\mathcal{A}(\hat{\phi})} > 0$. However, it could be the case that $\hat{\mu}_i = 0$ for some $i \notin \mathcal{A}(\hat{\phi})$, i.e. $i \in \mathcal{B}(\hat{\phi}) \setminus \mathcal{A}(\hat{\phi})$. Nonetheless, one can show that there is a set $\mathcal{N}$ of measure 0 for which $y \notin \mathcal{N}$ implies that $\mathcal{A}(\hat{\phi}(y)) = \mathcal{A}(\hat{\phi}(y'))$ and $\mathcal{B}(\hat{\phi}(y)) = \mathcal{B}(\hat{\phi}(y'))$ for all y' in a neighborhood of y. Lemma 9 of (Ryan) Tibshirani & Taylor (2012) proves this result for a nearly identical situation.

The fit $\widetilde{X}\hat{\phi}(y)$ is a piecewise affine function of y. Using Stein's formula for the degrees of freedom (as described in Ryan Tibshirani & Taylor 2012), we get that

$$df(\widetilde{X}\hat{\phi}) = E[\nabla \cdot \widetilde{X}\hat{\phi}(y)] = E[\text{tr}\{(\widetilde{X}P)(\widetilde{X}P)^+\}] = E[\text{rank}(\widetilde{X}P)],$$

where $P = P_{\text{null}(L)\cap\text{null}(D_{-\mathcal{A}(\hat{\phi})})}$.

D.2 Proof of bound on estimate (Property 5).

We bound this by an estimate that is more interpretable: $\text{rank}(\widetilde{X}P) \leq \text{rank}(P) = \text{nullity}\begin{pmatrix} L \\ D_{-\mathcal{A}(\hat{\phi})} \end{pmatrix}$.

Clearly, R2–R7 are linearly independent rows. Thus, the rank of $\widetilde{D}$ is at least $|\mathcal{Z}(\beta^+)| + |\mathcal{Z}(\beta^-)| + |\mathcal{Z}(\Theta^+)| + |\mathcal{Z}(\Theta^-)| + 2p$. Now, an R1 row is linearly independent of R2–R8 precisely when $j \in \mathcal{T}$ has $j \in \mathcal{Z}(\beta^+)\Delta\mathcal{Z}(\beta^-)$. To see this, note that if $j \in \mathcal{Z}(\beta^+) \setminus \mathcal{Z}(\beta^-)$, then R1 is certainly linearly independent of R2–R8 and likewise for $j \in \mathcal{Z}(\beta^-) \setminus \mathcal{Z}(\beta^+)$; however if $j \in \mathcal{Z}(\beta^+) \cap \mathcal{Z}(\beta^-)$ then $jk \in \mathcal{Z}(\Theta^+) \cap \mathcal{Z}(\Theta^-)$ for all $k \in \{1, \ldots, p\} \setminus \{j\}$ and therefore this row of R1 lies in the span of R3–R7. Thus, this means there are $|\mathcal{T} \setminus (\mathcal{Z}(\beta^+)\Delta\mathcal{Z}(\beta^-))|$ additional linearly independent rows. Finally, we consider R8. Clearly, R8 lies in the span of R4–R5 for $jk \in \mathcal{Z}(\Theta^+) \cap \mathcal{Z}(\Theta^-)$ since $jk \in \mathcal{Z}(\Theta^+) \implies kj \in \mathcal{Z}(\Theta^+)$ at a solution. But if $jk \in \mathcal{P}(\Theta^+) \cup \mathcal{P}(\Theta^-)$, then it is linearly independent of R1–R8. Therefore, R8 adds $|\mathcal{P}(\Theta^+)|/2 + |\mathcal{P}(\Theta^-)|/2$ to the rank where we have used that $\mathcal{P}(\Theta^+) \cap \mathcal{P}(\Theta^-) = \emptyset$ at a solution (since $\lambda_2 > 0$) and recalling that $j < k$ for the rows of R8. In summary, we have shown that the row-rank is

$$|\mathcal{Z}(\beta^+)| + |\mathcal{Z}(\beta^-)| + |\mathcal{Z}(\Theta^+)| + |\mathcal{Z}(\Theta^-)| + 2p + |\mathcal{T} \setminus (\mathcal{Z}(\beta^+)\Delta\mathcal{Z}(\beta^-))| + |\mathcal{P}(\Theta^+)|/2 + |\mathcal{P}(\Theta^-)|/2.$$

Recalling that there are $2p + 2p^2$ columns, we get that

$$\begin{aligned}
&\text{nullity}\begin{pmatrix} L \\ D_{-\mathcal{A}(\hat{\phi})} \end{pmatrix} = (2p + 2p^2) - \text{rank}\begin{pmatrix} L \\ D_{-\mathcal{A}(\hat{\phi})} \end{pmatrix} \\
&= (p - |\mathcal{Z}(\beta^+)|) + (p - |\mathcal{Z}(\beta^-)|) + (p(p-1) - |\mathcal{Z}(\Theta^+)|) + (p(p-1) - |\mathcal{Z}(\Theta^-)|) \\
&\quad - |\mathcal{T} \setminus (\mathcal{Z}(\beta^+)\Delta\mathcal{Z}(\beta^-))| - |\mathcal{P}(\Theta^+)|/2 - |\mathcal{P}(\Theta^-)|/2 \\
&= |\mathcal{P}(\beta^+)| + |\mathcal{P}(\beta^-)| + |\mathcal{P}(\Theta^+)|/2 + |\mathcal{P}(\Theta^-)|/2 - |\mathcal{T} \setminus (\mathcal{P}(\beta^+)\Delta\mathcal{P}(\beta^-))|
\end{aligned}$$

E Solving the prox function.

The Lagrangian of (11) is given by

$$\begin{aligned}
L(\beta^\pm, \theta; \alpha, \gamma^\pm) = &\frac{1}{2t}(\beta^+ - \tilde{\beta}^+)^2 + \frac{1}{2t}(\beta^- - \tilde{\beta}^-)^2 + \frac{1}{2t}\|\theta - \tilde{\theta}\|^2 + \left(\frac{\lambda}{2} + \alpha\right)\|\theta\|_1 \\
&- (\gamma^+ + \alpha)\beta^+ - (\gamma^- + \alpha)\beta^-
\end{aligned}$$

where α is the dual variable corresponding to the hierarchy constraints and $\gamma^\pm$ are the dual variables corresponding to the non-negativity constraints. For notational convenience, we have written θ for Θ_j and

we have dropped the subscripts on $\beta^\pm$. The KKT conditions are

$$\begin{aligned}
(\hat\beta^\pm - \tilde\beta^\pm)/t - \hat\gamma^\pm - \hat\alpha = 0 \qquad & (\hat\theta - \tilde\theta)/t + (\lambda/2 + \hat\alpha)u = 0 \\
\hat\beta^\pm \hat\gamma^\pm = 0 \qquad & \hat\alpha(\|\hat\theta\|_1 - \hat\beta^+ - \hat\beta^-) = 0 \\
\hat\beta^\pm \geq 0, \quad \|\hat\theta\|_1 \leq \hat\beta^+ + \hat\beta^- \qquad & \hat\alpha \geq 0, \quad \hat\gamma^\pm \geq 0,
\end{aligned}$$

where u_j is a subgradient of the absolute value function evaluated at $\hat\theta_j$. The three conditions involving $\gamma^\pm$ implies that $\hat\beta^\pm = [\tilde\beta^\pm + t\hat\alpha]_+$. The stationarity condition involving $\hat\theta$ implies that $\hat\theta = S(\tilde\theta,\ t(\lambda/2 + \hat\alpha))$. Now, define $f(\alpha) = \|S(\tilde\theta,\ t(\lambda/2+\alpha))\|_1 - [\tilde\beta^+ + t\alpha]_+ - [\tilde\beta^- + t\alpha]_+$. The remaining KKT conditions involve $\hat\alpha$ alone: $\hat\alpha f(\hat\alpha) = 0,\ f(\hat\alpha) \leq 0,\ \hat\alpha \geq 0$. Observing that f is non-increasing in α and piecewise linear suggests finding $\hat\alpha$ as done in Algorithm 3.

Algorithm 3 `ONEROW`: Solve (11) via dual.

Inputs: $\tilde\beta_j^+, \tilde\beta_j^- \in \mathbb{R},\ \widetilde\Theta_j \in \mathbb{R}^{p-1},\ \lambda \geq 0$.

1. Find $\hat\alpha$. Define $f(\alpha) = \|S(\widetilde\Theta_j,\ t(\lambda/2+\alpha))\|_1 - [\tilde\beta_j^+ + t\alpha]_+ - [\tilde\beta_j^- + t\alpha]_+$.
   (a) If $f(0) \leq 0$, take $\hat\alpha = 0$ and go to step 2.
   (b) Form knot the set $\mathcal{P} = \{|\widetilde\Theta_{jk}|/t - \lambda/2\}_{k=1}^p \cup \{-\tilde\beta^\pm/t\}$, and let $\mathcal{P}^+ = \mathcal{P} \cap [0,\infty)$.
   (c) Evaluate $f(p)$ for $p \in \mathcal{P}^+$.
   (d) If $f(p) = 0$ for some $p \in \mathcal{P}^+$, take $\hat\alpha = p$ and go to step 2.
   (e) Find adjacent knots, $p_1, p_2 \in \mathcal{P}^+$, such that $f(p_1) > 0 > f(p_2)$. Take
   $$\hat\alpha = -f(p_1)[f(p_2) - f(p_1)]/(p_2 - p_1).$$
2. Return $\widehat\Theta_j = \mathcal{S}[\widetilde\Theta_j,\ t(\lambda/2 + \hat\alpha)]$ and $\hat\beta_j^\pm = [\tilde\beta_j^\pm + t\hat\alpha]_+$.

F Solving the logistic regression problem.

For notational simplicity, in this section we use $\widetilde X$ and ϕ to denote the full data matrix and parameter combining both main effects and interactions. The binomial negative log-likelihood is

$$\ell(\beta_0, \phi) = -\sum_{i=1}^n [y_i \log p_i + (1-y_i)\log(1-p_i)]$$

where $p_i = p_i(\beta_0, \phi) = 1/(1 + e^{-\beta_0 - \tilde x_i^T \phi})$. Now,

$$\frac{\partial \ell(\beta_0,\phi)}{\partial \beta_0} = -1^T(y-p) \qquad\qquad \nabla_\phi \ell(\beta_0,\phi) = -\widetilde X^T(y-p).$$

Thus, to solve $\min_{\beta_0,\phi} \ell(\beta_0,\phi) + h(\phi)$, we can use generalized gradient descent, which iteratively solves

$$(\hat\beta_0^{(k)}, \hat\phi^{(k)}) = \arg\min_{\beta_0,\phi} \frac{1}{2t}\left\| \begin{pmatrix} \beta_0 \\ \phi \end{pmatrix} - \left[\begin{pmatrix} \hat\beta_0^{(k-1)} \\ \hat\phi^{(k-1)} \end{pmatrix} + t \begin{pmatrix} 1^T[y - p(\hat\beta_0^{(k-1)}, \hat\phi^{(k-1)})] \\ \widetilde X^T[y - p(\hat\beta_0^{(k-1)}, \hat\phi^{(k-1)})] \end{pmatrix} \right] \right\|^2 + h(\phi)$$

This separates into two parts:

$$\begin{aligned}
\hat\beta_0^{(k)} &= \hat\beta_0^{(k-1)} + t1^T[y - p(\hat\beta_0^{(k-1)}, \hat\phi^{(k-1)})] \\
\hat\phi^{(k)} &= \mathrm{Prox}_{2t\cdot h}\left(\hat\phi^{(k-1)} + t\widetilde X^T[y - p(\hat\beta_0^{(k-1)}, \hat\phi^{(k-1)})]\right),
\end{aligned}$$

where $\text{Prox}_{2t\cdot h}$ refers to the minimizer of (11). Looking at Algorithm 1, we see that this algorithm is identical except that for each k we update the estimate of the intercept and that we compute the residual as $y - p(\hat{\beta}_0, \hat{\phi})$. The "difficult" part of the computation is identical!

G ADMM for Strong Hierarchical Lasso.

The ADMM algorithm has three parts:

1. Update $(\beta_0,\ \beta^{\pm},\ \Theta)$ by solving

$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta^{\pm}\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}}\quad & q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1 \\ & + \text{tr}[U(\Theta - \widehat{\Omega})] + (\rho/2)\|\Theta - \widehat{\Omega}\|_F^2 \\ & \text{s.t.}\quad \beta_j^+ \geq 0, \beta_j^- \geq 0 \text{ for } j = 1, \ldots, p.\end{aligned}$$

   As with Algorithm 1, we may apply generalized gradient descent and `ONEROW` to solve this, but replacing the argument $\widetilde{\Theta}_j$ of `ONEROW` with $\delta\widehat{\Theta}_j^{(k-1)} - tZ_{(j,\cdot)}^T\hat{r}^{(k-1)} + \rho(\widehat{\Theta}_j^{(k-1)} - \widehat{\Omega}) + U$.

2. Update Ω by solving

$$\underset{\Omega\in\mathbb{R}^{p\times p}}{\text{Minimize}}\quad \text{tr}[U(\widehat{\Theta} - \Omega)] + (\rho/2)\|\widehat{\Theta} - \Omega\|_F^2 \quad \text{s.t.}\quad \Omega = \Omega^T.$$

   This has the analytic solution $\widehat{\Omega} \leftarrow \frac{1}{2}(\widehat{\Theta} + \widehat{\Theta}^T) + \frac{1}{2\rho}(U + U^T)$.

3. Update $U \leftarrow U + \rho(\widehat{\Theta} - \widehat{\Omega})$:

Algorithm 2 in the paper gives the full algorithm.

References

Agresti, A. (2002), *Categorical Data Analysis, Second Edition*, Wiley Series in Probability and Statistics, Wiley-Interscience.

Bach, F. (2011), 'Optimization with sparsity-inducing penalties', *Foundations and Trends® in Machine Learning* **4**(1), 1–106.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G. et al. (2012), 'Structured sparsity through convex optimization', *Statistical Science* .

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Bickel, P., Ritov, Y. & Tsybakov, A. (2008), 'Hierarchical selection of variables in sparse high-dimensional regression', *J. Roy. Statist. Soc. Ser. B* .

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–124.

Breiman, L. (1995), 'Best subset regression using the nonnegative garrote', *Technometrics* **37**(4), 373–384.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Chipman, H. (1996), 'Bayesian variable selection with related predictors', *Canadian Journal of Statistics* **24**(1), 17–36.

Choi, N., Li, W. & Zhu, J. (2010), 'Variable selection with the strong heredity constraint and its oracle property', *Journal of the American Statistical Association* **105**(489), 354–364.

Cox, D. (1984), 'Interaction', *International Statistical Review/Revue Internationale de Statistique* pp. 1–24.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *J. Amer. Statist. Assoc.* **81**, 461–70.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* (2), 407–499.

Forina, M., Armanino, C., Lanteri, S. & Tiscornia, E. (1983), 'Classification of olive oils from their fatty acid composition', *Food Research and Data Analysis* pp. 189–214.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics* **19**(1), 1–141.

Friedman, J. H., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

George, E. & McCulloch, R. (1993), 'Variable selection via gibbs sampling', *J. Amer. Statist. Assoc.* **88**, 884–889.

Hamada, M. & Wu, C. (1992), 'Analysis of designed experiments with complex aliasing', *Journal of Quality Technology* **24**, 130–137.

Jenatton, R., Audibert, J. & Bach, F. (2011), 'Structured variable selection with sparsity-inducing norms', *The Journal of Machine Learning Research* **12**, 2777–2824.

Jenatton, R., Mairal, J., Obozinski, G. & Bach, F. (2010), Proximal methods for sparse hierarchical dictionary learning, *in* 'Proceedings of the International Conference on Machine Learning (ICML)'.

McCullagh, P. & Nelder, J. (1983), *Generalized linear models*, Chapman & Hall/CRC.

Nardi, Y. & Rinaldo, A. (2007), 'The log-linear group lasso estimator and its asymptotic properties', *Arxiv preprint arXiv:0709.3526* .

Nelder, J. A. (1977), 'A reformulation of linear models', *Journal of the Royal Statistical Society. Series A (General)* pp. 48–77.

Nelder, J. A. (1997), 'Letters to the editors: Functional marginality is important', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 281–286.

Obozinski, G., Jacob, L. & Vert, J. (2011), 'Group lasso with overlaps: The latent group lasso approach', *arXiv preprint arXiv:1110.0413* .

Park, M. & Hastie, T. (2008), 'Penalized logistic regression for detecting gene interactions', *Biostatistics* **9**(1), 30–50.

Peixoto, J. (1987), 'Hierarchical variable selection in polynomial regression models', *American Statistician* pp. 311–313.

Radchenko, P. & James, G. M. (2010), 'Variable selection using adaptive nonlinear interaction structures in high dimensions', *Journal of the American Statistical Association* **105**(492), 1541–1553.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. & Shafer, R. (2006), 'Genotypic predictors of human immunodeficiency virus type 1 drug resistance', *Proceedings of the National Academy of Sciences* **103**(46), 17355.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**, 1131–1151.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. Royal. Stat. Soc. B.* **58**, 267–288.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232.

Tibshirani, R. & Taylor, J. (2011), 'The Solution Path of the Generalized Lasso', *Annals of Statistics* **39**(3), 1335–1371.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of optimization theory and applications* **109**(3), 475–494.

Turlach, B. (2004), 'Discussion of "least angle regression"', *Annals of Statistics* (2), 481–490.

Wu, J., Devlin, B., Ringquist, S., Trucco, M. & Roeder, K. (2010), 'Screen and clean: a tool for identifying interactions in genome-wide association studies', *Genetic epidemiology* **34**(3), 275–285.

Yuan, M., Joseph, V. & Lin, Y. (2007), 'An efficient variable selection approach for analyzing designed experiments', *Technometrics* **49**(4), 430–439.

Yuan, M., Joseph, V. & Zou, H. (2009), 'Structured variable selection and estimation', *The Annals of Applied Statistics* **3**(4), 1738–1757.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhao, P., Rocha, G. & Yu, B. (2009), 'The composite absolute penalties family for grouped and hierarchical variable selection', *The Annals of Statistics* **37**(6A), 3468–3497.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. Royal. Stat. Soc. B.* **67**, 301–320.

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the "degrees of freedom" of the lasso', *The Annals of Statistics* **35**(5), 2173–2192.